THE UNIVERSITY OF
WARWICK

University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

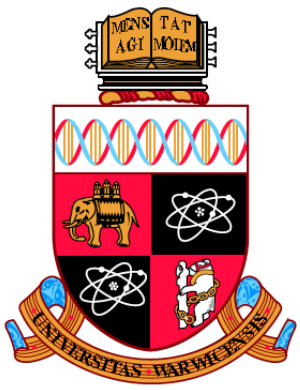
A Thesis Submitted for the Degree of PhD at the University of Warwick

http://go.warwick.ac.uk/wrap/77353

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



Radiomics in Paediatric Neuro-Oncology: MRI Textural Features as Diagnostic and Prognostic Biomarkers

by

Ahmed E. Fetit

A thesis submitted for the degree of

Doctor of Philosophy in Engineering

Institute of Digital Healthcare, University of Warwick

July 2015

Contents

Acknowledgements **xvii**

Declerations **xix**

Publications **xx**

Abstract **xxi**

Abbreviations **xxii**

1 Introduction **1**

1.1 Motivation 2

1.2 Aim and Objectives 6

1.3 Contributions to Knowledge 7

1.4 Thesis Structure 8

2 Background on MR Imaging of Paediatric Brain Tumours **12**

2.1 Introduction 13

2.2 Background on MR Imaging 13

2.2.1 Nuclear Magnetic Resonance 13

2.2.2 Radio Frequency Pulses 16

2.2.3 Rotating Frame of Reference 17

2.2.4 Longitudinal (T1) Relaxation 19
2.2.5 Transverse (T2) Relaxation 20
2.2.6 Magnetic Field Gradients 23
2.2.7 Slice Selection 24
2.2.8 Frequency and Phase Encoding 26
2.2.9 K-Space 28
2.2.10 Imaging Planes 29
2.2.11 Image Contrast 31
2.3 MR Imaging Characteristics of Brain Tumours in Children 35
2.3.1 Basic Brain Anatomy 35
2.3.2 Paediatric Brain Tumours 36
2.3.3 Comparison to Adult Brain Tumours 43
2.4 Summary 44

3 Background on Machine Learning 45

3.1 Introduction 46
3.2 The Learning Problem 47
3.2.1 Aim of a Learning Task 47
3.2.2 Types of Learning 47
3.2.3 The Supervised Classification Framework 49
3.3 Feature Selection and Dimensionality Reduction 50
3.4 Classification Methods 52
3.4.1 The Bayes Classifier 52
3.4.2 k-Nearest Neighbours Classifier 53
3.4.3 Classification Tree 54
3.4.4 Logistic Regression 56
3.4.5 Artificial Neural Network 58

3.4.6 Support Vector Machines 61
3.5 Model Validation and Evaluation 68
3.5.1 Measures of Classification Performance 68
3.5.2 Model Validation Schemes 72
3.5.3 The Problem of Over-fitting 72
3.5.4 The Problem of Class-Imbalance 73
3.6 Summary 74

4 Texture Analysis of MR Images: Theory and State-of-the-Art 75
4.1 Introduction 76
4.2 Background on Texture Analysis 76
4.2.1 Introduction 76
4.2.2 Common Statistical TA Methods 78
4.2.3 Three-dimensional Texture Analysis 93
4.2.4 Practical Limitations of MRI Texture Analysis 96
4.3 Review of the Current State-of-the-Art 97
4.3.1 Applications of TA in Diagnostic Classification of Paediatric Brain Tumours 97
4.3.2 Applications of TA in Diagnostic Classification of Adult Brain Tumours 100
4.3.3 Other Diagnostic Applications of MRI TA 105
4.3.4 TA for Estimating Survival Prognosis 105
4.4 Summary 106

5 A Single Centre Study on 3D TA 108
5.1 Introduction 109
5.2 Materials and Methods 110

5.2.1 Cohort Details and Image Acquisition . . . 110
5.2.2 Image Pre-processing . . . 111
5.2.3 Textural Features Extraction . . . 113
5.2.4 Feature Selection and Analysis . . . 113
5.2.5 Statistical Analysis . . . 117
5.2.6 Obtaining a Radiological Review Benchmark . . . 118
5.3 Results . . . 119
5.3.1 Top ranked 2D features . . . 119
5.3.2 Top ranked 3D features . . . 119
5.3.3 Classification Results and Statistical Findings . . . 124
5.3.4 PCA-based Results . . . 134
5.3.5 Radiological Reporting Benchmark . . . 137
5.4 Discussion . . . 138
5.5 Study Limitations and Future Work . . . 142
5.6 Conclusion . . . 142

6 A Multicentre Investigation on the Transferability of TA 144
6.1 Introduction . . . 145
6.2 Materials and Methods . . . 146
6.2.1 Cohort Details and Image Acquisition . . . 146
6.2.2 Image Pre-processing . . . 146
6.2.3 Extraction of Textural Features . . . 147
6.2.4 Feature Selection . . . 147
6.2.5 Classification Model . . . 147
6.2.6 Model Validation . . . 147
6.2.7 Addressing the Class Imbalance Problem . . . 150
6.3 Results . . . 151

6.3.1 Results from Pairwise Testing on Unseen Data 151
6.3.2 LOOCV Results 158
6.3.3 LOOCV Results After Minority-Oversampling 160
6.4 Discussion 162
6.5 Study Limitations and Future Work 165
6.6 Conclusions 165

7 Predicting Survival in Paediatric Medulloblastoma 166
7.1 Introduction 167
7.2 Materials and Methods 168
7.2.1 Cohort Details and Image Acquisition 168
7.2.2 Image Pre-processing 168
7.2.3 Extraction of Textural Features 169
7.2.4 Identifying Textural Features with Potential Prognostic Value 169
7.2.5 Statistical Methods 171
7.3 Results 171
7.4 Discussion 178
7.5 Study Limitations and Future Work 179
7.6 Conclusion 181

8 Summary, Conclusion and Recommendations for Future Work 182
8.1 Summary 183
8.2 Conclusion 185
8.3 Recommendations for Future Work 186

A Preliminary Study 189
A.1 Introduction 190
A.2 Materials and Methods 190

A.2.1 Clinical Materials 190
A.2.2 Image Pre-processing 191
A.2.3 Textural Features Extraction 193
A.2.4 Feature Selection and Supervised Learning 193
A.3 Preliminary Results 194
A.4 Discussion 198
A.5 Conclusion 199

B Non-Statistical TA Techniques **201**

References **205**

Colophon **216**

List of Figures

1.1 A figure showing three T2-weighted MR images of (a)Medulloblastoma (b) Pilocytic Astrocytoma and (c) Ependymoma, the three most common brain tumours occuring in childhood. 2

1.2 Example biopsy micrographs for (a) Medulloblastoma (b) Pilocytic Astrocytoma and (c) Ependymoma. 3

2.1 A diagram showing spin orientations when placed in an external magnetic field $\vec{B_0}$. 14

2.2 A diagram showing how spins precess around the axis of an applied magnetic field $\vec{B_0}$. 14

2.3 A vector model of NMR representing spin orientations and magnetisation vector $\vec{M_0}$. 15

2.4 A diagram showing a 90 degree RF pulse being used to manipulate $\vec{M_0}$ into the XY plane. 16

2.5 A diagram depicting how, after the 90 degree excitation pulse, the magnetisation vector gradually returns to its equilibrium state. . . . 17

2.6 A plot showing how the free induction decay (FID) signal decays with time as the system returns to equilibrium. 17

2.7 A diagram showing the pulse sequence timings for an inversion recovery experiment to measure longitudinal relaxation (T1). 19

2.8 A plot depicting the relationship between the recovery of longitudinal magnetisation and time delay in an inversion recovery experiment. 20
2.9 A diagram showing the pulse sequence timings for a spin echo (SE) experiment. The time to echo formation is referred to as echo time (TE), while the time between successive excitations is referred to as repetition time (TR). 21
2.10 A diagram showing the de-phasing and refocusing of the magnetisation vector during a spin echo pulse sequence. 21
2.11 A diagram showing the formation of spin echoes by multiple 180 degree pulses (the pulse sequence and the resulting echo signals). . 22
2.12 A diagram showing three test tubes filled with water and how the use of an external magnetic field affects the resonant frequencies across the tubes 23
2.13 A diagram showing slice selection being carried out on a patient using a magnetic field gradient and an RF pulse. 25
2.14 A diagram showing the time and frequency excitation profiles of the desired RF excitation pulse.. 25
2.15 A diagram illustrating the use of a field gradient for carrying out phase encoding. 27
2.16 A diagram showing how the overall k-space trajectory may be a set of horizontal traces stacked one above the other. 28
2.17 A diagram showing how 2D Fourier Transform can change MR data of a human brain in k-space to image space. 29
2.18 A diagram showing a number of brain MR image slices in three different planes: (a) axial, (b) sagittal and (c) coronal. 30

2.19 (a) T1 and (b) T2-weighted brain MR images demonstrating how image contrast can be manipulated using T1 and T2 relaxation times. 32
2.20 (a) T1 and (b) T2-weighted brain MR images demonstrating how image contrast can be manipulated using T1 and T2 relaxation times. 33
2.21 MR scan of a child showing brain structure. 36
2.22 (a) T1- and (b) T2-weighted MR images of a child diagnosed with medulloblastoma. 38
2.23 (a) T1- and (b) T2-weighted MR images of a child diagnosed with pilocytic astrocytoma. 40
2.24 (a) T1- and (b) T2-weighted MR images of a child diagnosed with ependymoma. Images are shown in the axial and coronal planes. . . 42

3.1 A figure showing (a)the idealisation of a perceptron. Each activity is multiplied by a weight and the weighted inputs are then added. The output activity is computed using an activation function (b) an ANN consisting of three layers that are fully connected. 59
3.2 A graph showing a one-dimensional hyperplane 61
3.3 A graph showing a number of data points that fall into one of two possible classes. 62
3.4 A plot illustrating a maximal margin hyperplane separating data points that belong to two classes. 63
3.5 A plot illustrating the construction of a soft margin that allows two data points to be on the incorrect side of the hyperplane and margin, but performs well when separating the rest of the data points. 64

3.6 A plot illustrating two classes of data that are not linearly-separable. 65
3.7 A plot illustrating thow polynomial (left) and radial (right) kernels perform on data. 66
3.8 A diagram showing an example confusion matrix. 69
3.9 A diagram showing an example receiver operator characteristics (ROC) curve. 70

4.1 Different types of image textures that human vision processes on a daily basis. 76
4.2 Four regions of interest (ROIs) and their corresponding histograms extracted from an axial T1-weighted MR image. 79
4.3 (a) An axial T1-weighted MR image and (b) its corresponding gradient image. 81
4.4 A grey-level image showing a hypothetical pixel neighbourhood. . . 82
4.5 (a)Illustration of the pixel relationships considered by the Grey-Level Co-Occurence Matrix (GLCM) technique.(b)A hypothetical image and its corresponding GLCM, assuming a horizontal direction of analysis and a 1-pixel distance 84
4.6 Two T1-weighted MR images of a (a) medulloblastoma and a (b) pilocytic astrocytoma. 87
4.7 (a)Illustration of the pixel relationships considered by the Grey-Level Run-Length Matrix (GLRLM) technique.(b)A hypothetical image and its corresponding GLRLM, assuming a horizontal direction of analysis and a 1-pixel distance 89
4.8 Two T1-weighted MR images of a (a) medulloblastoma and a (b) pilocytic astrocytoma. 92

4.9 An illustration of the spatial relationship between voxels on a single two-dimensional image slice (left) and a three-dimensional multi-slice volume (right). 93
4.10 Multiple axial T2-weighted MR slices for one child diagnosed with medulloblastoma. 94
4.11 A figure illustrating θ and ϕ, which are used to spatially characterise directions of analysis in 3D GLCMs [111]. 94

5.1 A figure showing semi-automatic segmentation of a tumour region of interest using the Snake GVF algorithm. 111
5.2 Distance maps based on Pearson Correlation metric, measured for variations of three features: T2 Angular Second Moment, Sum of Squares and Sum Variance. 122
5.3 Bar charts summarising AUC values obtained with 2D and 3D features on LOOCV for each of the classifiers. 125
5.4 Shows a scatter plot of two 3D features used to train SVM classifier, namely T1 Sum of Squares (0,0,3) and T2 Sum Average (0,0,2). . . 127
5.5 A scatter plot of two 3D features used to train SVM classifier, namely T2 Sum Variance (0,1,0) and T2 Skewness. 128
5.6 A bar chart that summarises the obtained AUC values when T1 and T2-weighted features were tested independently using LOOCV on the 2D and 3D datasets. 131
5.7 A bar plot of AUC results obtained with the PCA-based pipeline. . 135
5.8 A scatter plot of PC1 vs. PC2 using 2D features 136
5.9 A scatter plot of PC1 vs. PC2 using 3D features 136
5.10 A bar plot summarising probabilities assigned to each individual diagnosis by the neural network classifier during LOOCV. 141

6.1 A flowchart showing methodological overview of the multicentre experimental set up. 149
6.2 Bar chart showing optimal AUC values obtained through pairwise testing for multicentre classification. 156
6.3 Bar chart showing number of features that were needed for optimal AUC. 157
6.4 ROC curves depicting SVM classifier performance using the LOOCV scheme. All 134 samples obtained from three hospitals were used for the analysis. 159
6.5 ROC curves depicting SVM classifier performance using the LOOCV scheme, after SMOTE was used to generate 27 synthetic ependymoma samples. 161

7.1 Kaplan-Meier survival curves for five of the fifteen features identified to be of prognostic value: *T2 Sum Variance* (1,-1,0),(1,0,0), (2,-2,0),(2,0,0),(0,0,3). 173
7.2 Kaplan-Meier survival curves for two of the fifteen features identified to be of prognostic value: *T2 Sum of Squares* (1,1,0) ,(0,0,3). 174
7.3 Kaplan-Meier survival curves for four of the fifteen features identified to be of prognostic value: *T2 Angular Second Moment* (2,-2,0), (0,2,0), (2,2,0), (3,0,0). 175
7.4 Kaplan-Meier survival curves for four of the fifteen features identified to be of prognostic value: *T2-weighted Angular Second Moment*(0,3,0),(3,3,0),(4,0,0),(4,4,0). 176
7.5 Four T2-weighted medulloblastoma ROIs and their corresponding feature maps. 180

A.1 A figure showing the placing of a 30x30 pixel region of interest on the tumour region of a T2-weighted image of a medulloblastoma, using the MaZda software. 192

A.2 A figure showing the receiver operator characteristics (ROC) curves for group (a): embryonal vs astrocytic. 195

A.3 A figure showing the receiver operator characteristics (ROC) curves for group (b): embryonal vs ependymal. 196

A.4 A figure showing the receiver operator characteristics (ROC) curves for group (c): ependymal vs astrocytic. 197

B.1 A hypothetical pixel neighbourhood showing a pixel s and its surrounding region (shaded in grey) where a casual AR model neighbourhood may be located. 203

B.2 An axial T2-weighted MR image (left) and its corresponding wavelet transform sub-bands (right). 204

List of Tables

4.1 Offsets describing 13 possible directions of analysis when computing 3D GLCMs and GLRLMs. d is the chosen distance of analysis, in number of pixels. 95

4.2 TA articles available in the literature that look into classifying childhood brain tumours from MR images. 100

4.3 A summary of TA articles available in the literature that look into classifying adult brain tumours from MR images. 104

5.1 A table summarising the TA methods used and their corresponding features. 113

5.2 A table showing a breakdown of the number of textural features for each dataset. 114

5.3 A table showing a summary of the T1 and T2-weighted 2D features chosen by entropy-MDL during the feature selection stage. 121

5.4 A table showing a summary of the T1 and T2-weighted 3D features chosen by entropy-MDL during the feature selection stage. 123

5.5 Summary of classification results obtained by leave-one-out cross-validation (LOOCV) on 2D and 3D textural features. 126

5.6 Contingency table constructed from 3D and 2D-trained SVM classifiers for performing McNemar's test. 129

5.7 Table summarising results obtained by McNemar's test to assess whether 3D and 2D trained classifiers showed significant differences in performance. 129
5.8 Confidence intervals for overall classification accuracies, obtained by a bootstrapping of samples 1000 times. 130
5.9 Summary of classification results obtained by stratified 10-fold cross-validation on 2D and 3D textural features. 133

6.1 A table summarising models and field strengths for the three centres. 146
6.2 Optimal AUC values obtained through pairwise testing for multi-centre classification. 152
6.3 A table listing the number of features that were needed to yield optimal AUC. 153
6.4 A table listing the optimal textural features identified in Tests 1 to 3. 154
6.5 A table listing the optimal textural features identified in Tests 4 to 6. 155
6.6 A table listing the results obtained when the feature-set, comprising data from all three hospitals (134 samples), was tested with an SVM classifier on LOOCV. 158
6.7 A table listing the classification results obtained with LOOCV, after SMOTE was applied to generate 27 synthetic EP samples. 160

7.1 Summary of the textural features identified to be of significant prognostic value by log-rank test. Note that all features were extracted from T2-weighted images. 172
7.2 A table summarising linear correlation coefficients obtained by applying Pearson's test on the 15 optimal features. 177

A.1 A table summarising the datasets included in this preliminary study. 190

A.2 A table summarising the classification accuracies and the corresponding AUC values obtained with each of the four classifiers on all three datasets. Model validation was carried out using random sampling, where the training/testing process was repeated 20 times. 198

Acknowledgements

Although only my name appears on the front page of this thesis, many people have helped me during the course of this research project. I owe my gratitude to all those who, knowingly or unknowingly, helped make this work possible.

Throughout this PhD, I had the fortune of having three people overseeing my work. Professor Theodoros Arvanitis gave me the opportunity to explore my ideas and curiosities in a very relaxed atmosphere; his advice and direction are much appreciated. Professor Andrew Peet gave me the exciting opportunity of being a member of the Brain Tumour Research Group and ensured the clinical applicability of this research; I am very grateful for all your support throughout the difficult times of trying to publish my first journal paper. Dr Jan Novak taught me how to think like a scientist and how to approach problems pragmatically; thank you for your patience and for your numerous suggestions on how to improve my academic writing.

I take this opportunity to also thank my examiners, Professor Thomas Nichols and Professor Paul Morgan, for their views, comments and stimulating discussion during the viva, which helped this thesis reach its final version.

I am very grateful to (soon-to-be Dr) Soroosh Afyouni for all the coffee breaks and for helping make the difficult daily commute to Warwick very enjoyable. Your enthusiasm and discussions on the Erdos-Renyi mixture model have indeed been very inspirational.

Thank you, to everyone at the University of Warwick who made me feel welcome when I first joined the Institute of Digital Healthcare, particularly Josh Elliott, Omar Khan, Chris Golby, Camille Maumet, Lorena Santamaria, Naeem Mohammed and Sallyann Edwards. Thank you, Huseyin Dereli, for all the discussions on entrepreneurship and mobile app ideas.

I am also very thankful to Sarah Lim Choi Keung and Lei Zao for being wonderful officemates, in both Birmingham and Warwick.

Thank you, to all members of the Brain Tumour Research Group at Birmingham Children's Hospital for allowing me to attend and present at the Friday group meetings. I must also thank everyone I met at the Gisbert Kapp building at the University of Birmingham, where I spent the first two years of my doctorate; in particular, Ahmed Zaidi and Natan Morar for their continuous encouragement.

Grateful acknowledgement is made to Warwick Manufacturing Group (University of Warwick) as well as the TI Group Scholarship (University of Birmingham) for partial financial support.

Monica Stef contributed enormously towards improving my writing skills; thank you for your continuous support throughout my research.

And ofcourse, I would like to thank my friends who I met in Birmingham, for always keeping me laughing.

Above all, I am most grateful to my parents and my sister Rana for their love, support and encouragement throughout my research. I owe you much more than what I could ever express in words.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below:

- Data was collected from three hospitals: Birmingham Children's Hospital, Nottingham University Hospital and Great Ormond Street Hospital. The staff of the radiology departments in these hospitals were responsible for collecting magnetic resonance imaging data. Histopathology was undertaken by the pathology department within each hospital.
- Multicentre ethical approval was given and parental consent was obtained. Study title: Functional Imaging of Tumours (CNS 2004 10); Research Ethics Committees reference: 04/MRE04/41; Sponsor reference: RG 09-028.
- The Children's Cancer and Leukaemia Group (CCLG) Functional Imaging Group provided a secure e-repository [4] from which the datasets were downloaded.
- The review of radiological reports from Birmingham Children's Hospital was done with the assistance of Professor Andrew Peet.
- Patient survival information was summarised by Dr. Simrandip Gill and Dr. Martin Wilson of Birmingham Children's Hospital, before being given to the author.

Publications

Journal publications based on this work:

- [**P01**] A.E. Fetit, J. Novak, A.C. Peet and T.N. Arvanitis, "Three-dimensional textural features of conventional MRI improve diagnostic classification of childhood brain tumours," *NMR in Biomedicine,* 2015: volume 28, issue 9: 1174-1184.

Conference proceedings based on this work:

- [**P02**] A.E. Fetit, J. Novak, D. Rodriguez, D.P. Auer, C.A. Clark, R.G. Grundy, T. Jaspan, A.C. Peet and T.N. Arvanitis, "MRI texture analysis in paediatric oncology: a preliminary study," *Studies in Health Technology and Informatics,* 2013; 190:169-71.

- [**P03**] A.E. Fetit, J. Novak, A.C. Peet and T.N. Arvanitis, "3D texture analysis of MR images to improve classification of paediatric brain tumours: a preliminary study," *Studies in Health Technology and Informatics,* 2014; 202:213-6.

- [**P04**] A.E. Fetit, J. Novak, D. Rodriguez, D.P. Auer, C.A. Clark, R.G. Grundy, T. Jaspan, A.C. Peet and T.N. Arvanitis "3D texture analysis of heterogeneous MRI data for diagnostic classification of childhood brain tumours" *Studies in Health Technology and Informatics,* 2015; 213:19-22.

  This publication was awarded the **"Best Student Paper Award"** by the 13th International Conference on Informatics, Management and Technology in Healthcare (ICIMTH), which was held in Athens, Greece, in July 2015.

Conference abstracts based on this work:

- [**P05**] A.E. Fetit, J. Novak, D. Rodriguez, D.P. Auer, C.A. Clark, R.G. Grundy, T. Jaspan, A.C. Peet and T.N. Arvanitis "3D texture analysis of heterogeneous MRI data for the characterisation of childhood brain tumours," *International Society of Magnetic Resonance in Medicine (ISMRM) Annual Scientific Meeting, Milan, Italy.* 2014: 4466.

- [**P06**] A.E. Fetit, J. Novak, S.K. Gill, M. Wilson, A.C. Peet and T.N. Arvanitis "3D textural features of conventional MRI predict survival in childhood medulloblastoma," *International Society of Magnetic Resonance in Medicine (ISMRM) Annual Scientific Meeting, Toronto, Canada.* 2015: 4494.

Abstract

Motivation: Brain and central nervous system tumours form the second most common group of cancers in children in the UK, accounting for 27% of all childhood cancers. Despite current advances in magnetic resonance imaging (MRI), non-invasive characterisation of paediatric brain tumours remains challenging. Radiomics, the high-throughput extraction and analysis of quantitative image features (e.g. texture), offers potential solutions for tumour characterisation and decision support.

Aim and Methods: In search for diagnostic and prognostic oncological markers, the aim of this thesis was to study the application of MRI texture analysis (TA) for the characterisation of paediatric brain tumours. To this end, single and multi-centre experiments were carried out, within a supervised classification framework, on clinical MR imaging datasets of common brain tumour types.

Results: TA of conventional MRI was successfully used for diagnostic classification of common paediatric brain tumours. A key contribution of this thesis was to provide evidence that diagnostic classification could be optimised by extending the analysis to include three-dimensional features obtained from multiple MR imaging slices. In addition to this, TA was shown to have a good cross-centre transferability, which is essential for long-term clinical adoption of the technique. Finally, fifteen textural features extracted from T2-weighted MRI were identified to be of significant prognostic value for paediatric medulloblastoma.

Conclusion: It was shown that MRI TA provides valuable quantifiable information that can supplement qualitative assessments conducted by radiologists, for the characterisation of paediatric brain tumours. TA can potentially have a large clinical impact, since MR imaging is routinely used in the brain cancer clinical work-flow worldwide, providing an opportunity to improve personalised healthcare and decision-support at low cost.

Keywords: Texture analysis, MRI, brain tumours, paediatrics, machine learning.

Abbreviations

AUC = Area under the ROC Curve

ANN = Artificial Neural Network

CV = Cross Validation

CCLG = Children's Cancer and Leukaemia Group

DICOM = Digital Imaging and Communication in Medicine

EP = Ependymoma

GLCM = Grey-Level Co-Occurrence Matrix

GLRLM = Grey-Level Run-Length Matrix

GVF = Gradient Vector Flow

kNN = k-Nearest Neighbour

LOOCV = Leave-One-Out Cross Validation

MRI = Magnetic Resonance Imaging

MDL = Minimum Descriptive Length

MB = Medulloblastoma

PA = Pilocytic Astrocytoma

PCA = Principal Component Analysis

ROC = Receiver Operating Characteristics

ROI = Region of Interest

SVM = Support Vector Machine

T1 = Longitudinal Relaxation

T2 = Transverse Relaxation

TR = Repetition Time

TE = Echo Time

TA = Texture Analysis

For mom, dad and Rana

Chapter 1

Introduction

In preparing this work, I am reminded of the inspiring comment from Galileo Galilei: *"measure what is measurable, and make measurable what is not so"*. This thesis aims to push the limit to what can be measured from magnetic resonance images, by capturing information that may be below human vision but of potential value to the clinically important area of paediatric oncology. To this end, the material presented in this chapter provides an introduction to the problem domain and outlines the aim, objectives and contributions of this work.

1.1 Motivation

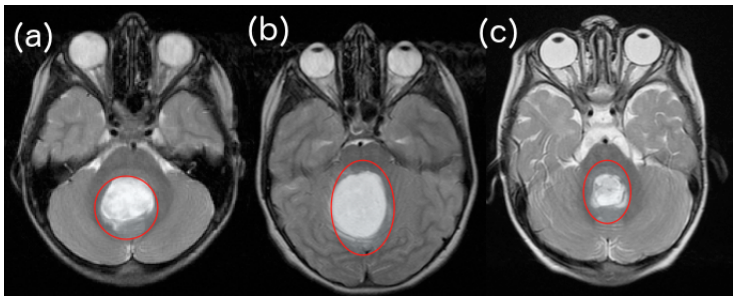


Figure 1.1: A figure showing three T2-weighted MR images of (a)Medulloblastoma (b) Pilocytic Astrocytoma and (c) Ependymoma, the three most common brain tumours occurring in childhood. Tumour regions are marked in red. Besides variation in size, the tumours do not show clear differences in visual appearance. Original images were obtained from the CCLG database [4].

Cancer is a leading cause of mortality in children, with the latest available statistics in the UK showing that between 2009 and 2011, an average of 1,574 children per year were diagnosed with cancer, of which 16% had died [74]. Brain and central nervous system (CNS) tumours form the second most common group of cancers in children, accounting for 27% of all childhood cancers [74]. In order to tailor surgery and drug-based therapy, a brain tumour must be classified as one of 37 types, as outlined by the World Health Organisation (WHO) [75], [84].

Magnetic resonance imaging (MRI) is the key imaging technique used for visualising and managing brain tumours [76], [77]. Initial characterisation of tumours from MRI scans is usually performed via radiologists′ visual assessment [78]. However, different brain tumour types do not always demonstrate clear differences in visual appearance [37]. Using only conventional MRI to provide a definite diagnosis could potentially lead to inaccurate results, so histopathological examination of biopsy samples are currently considered the gold standard for obtaining definite diagnoses [77].

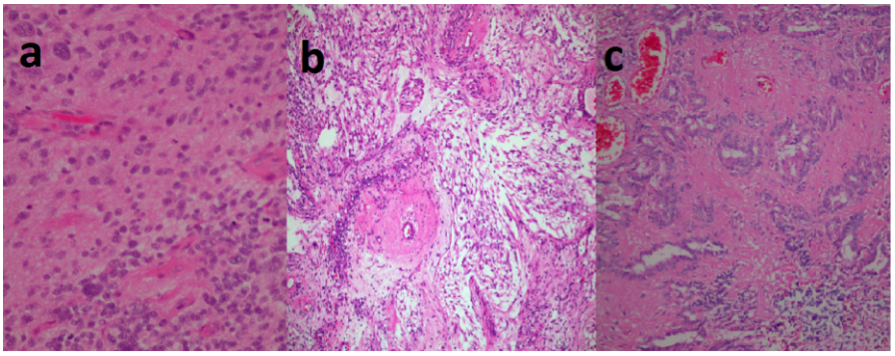


Figure 1.2: Example biopsy micrographs for (a) Medulloblastoma (b) Pilocytic Astrocytoma and (c) Ependymoma. One can see how different paediatric brain tumours demonstrate clear differences in visual characteristics on a microscopic scale. For example, medulloblastoma is characterised by its solid, well-circumscribed appearance, whereas pilocytic astrocytoma micrographs tend to show characteristic bipolar cells with long hair-like structures.

By inspecting Figure 1.1, one could see how different tumour types could demonstrate similar visual characteristics on MR images. On a microscopic scale, however, biopsies of different paediatric brain tumours demonstrate clear differences in visual characteristics. For example, medulloblastoma is characterised by its solid, well-circumscribed appearance, whereas pilocytic astrocytoma biopsy micrographs tend to show characteristic bipolar cells with long hair-like structures, as can be seen in Figure 1.2. It is likely that such microscopic patterns translate to small dissimilarities and subtle visual differences on MR images. The avail-

ability of non-invasive diagnostic aids that could capture such patterns from MR images would be very beneficial. Benefits of such tools would include reducing surgical procedures, improving surgery and therapy planning and the possibility to support more informed discussions with the patient's family.

The emerging field of *radiomics* provides a potential solution for non-invasive tumour characterisation by converting medical images into mineable data, through the extraction of a large number of quantitative imaging features [83], [88]. When developing quantitative medical image analysis techniques, it is usual to consider attributes which radiologists explicitly or implicitly use in their assessment of a specified tissue appearance. Intensity, morphology and texture are common examples of such important image attributes [39], [79]. Image texture can be defined as the spatial variation of pixel intensities within an image [18], and is known to be particularly sensitive for the assessment of pathology [39]. Visual assessment of texture is, however, particularly subjective. Additionally, it is known that human observers possess limited sensitivity to textural patterns, whereas computational texture analysis (TA) techniques can be significantly more sensitive to changes [39], [79].

MRI TA has been recently used with success, within machine learning frameworks, to discriminate between childhood brain tumours, as reported by Rodriguez Guiterrez et al [36] and Orphanidou-Vlachou et al [37]. There exists, therefore, considerable motivation for further research into maximising the value of computational TA as a predictive biomarker in paediatric oncology, and the establishment of its role in clinical practice.

Whilst most of the MRI TA experiments reported in the cancer literature focused on the analysis of textural features derived from a 2D image slice, there have been recent efforts to extend analysis to multiple MR image slices. The

processing of multi-slice volumetric features may offer additional information that will improve classification performance [26], [33], [34]. Using only one 2D slice, as representative of the entire tumour, might not be sufficient for building a reliable classification model, as capturing any heterogeneities present across the tumour volume would not be possible. In addition to this, 3D TA has the advantage of capturing inter-slice features that are completely ignored in the traditional 2D approach. When characterising brain MRI scans, radiologists base their decisions on multi-slice imaging information and do not analyse individual slices in isolation. None of the work available in the *childhood brain cancer* literature, however, looked into the use of 3D TA of MR images, at the time of writing and to the best of the author's knowledge.

The reported diagnostic success of MRI TA raises an interesting question: *If textural features could capture powerful patterns that aid the diagnosis of tumours, can they also be used to predict patients' survival prognosis?* Following diagnosis, determination of prognosis is an important step in brain tumour management, with implications that determine treatment options. Therefore, accurate non-invasive predictors of prognosis have the potential to advance clinical management of patients for therapy and the possibility to support more informed discussions with the patient's family.

This thesis makes use of MRI datasets obtained from ongoing studies at Birmingham Children′s Hospital. In addition to this, some aspects of this work make use of multicenter datasets obtained from Great Ormond Street Hospital and University Hospital Nottingham[1].

[1]All images were anonymised and held at a secure e-repository provided by Children′ Cancer and Leukaemia Group (CCLG) Functional Imaging Group [4]. Approval was obtained from the research ethics committee and informed consent was taken from patients′ guardians.

1.2 Aim and Objectives

On the basis of the previous section's discussion, the aim of this thesis is to study the application of MRI TA for the characterisation of childhood brain tumours. The problem of tumour characterisation can be divided into two parts: diagnosis and prognosis. To this end, three specific objectives are defined as follows.

The **first objective** of this thesis is to carry out a practical investigation using clinical datasets in order to assess the efficacy of MRI TA in classifying childhood brain tumour types. To ensure long-term clinical adoption of TA as a diagnostic tool, maximising its classification performance is crucial. Hence, and in light of recent efforts towards carrying out 3D TA in the adult literature, the investigation would require rigorous analysis into whether 3D TA could capture more discriminative tumour patterns than the traditional 2D approach.

Despite the positive results reported in the adult and childhood MRI literature, TA has not yet found its way into routine clinical practice. This is perhaps due to the sensitivity of TA to variations in MR acquisition parameters, which may impede the transfer of results across various imaging centres. Therefore, the **second objective** of this thesis is to determine, on a multicentre level, the efficacy and transferability of 3D TA for diagnostic classification of childhood brain tumours.

The **final objective** is to study the application of TA in the problem of predicting patients′ survival prognosis. If such application could be proven possible, this will have potential long-term benefits of supporting the determination of treatment options and advancing clinical management of patients for therapy.

1.3 Contributions to Knowledge

The thesis offers three novel contributions to the area of non-invasive, computational characterisation of childhood brain tumours from textural features of MR images.

The first is the application of a 3D TA framework specifically designed to classify the three most frequently occurring types of brain tumours in children: medulloblastoma, pilocytic astrocytoma and ependymoma. This was done through experimental analysis that used clinical MRI datasets obtained from currently ongoing studies at Birmingham Children′s Hospital. The analysis included a rigorous comparison of 3D TA to the traditional 2D approach, which is the current state-of-the-art in the paediatric literature.

Building on the first contribution of the work, the second contribution is the multicenter investigation of the efficacy and transferability of 3D MRI TA using datasets obtained from three different hospitals: Birmingham Children′s Hospital, Nottingham University Hospital and Great Ormond Street Hospital. An essential outcome of this study is that, despite the variations in textural information among MR images from different centres, feature-sets acquired from one centre can be used for successful tumour classification in unseen data from other centres. This analysis also included an investigation on the nature of features that are most likely to train classifiers, which can generalise well with the 3D textural data. Additionally, the issue of class imbalance, which arises because some tumour types do not occur as frequently as others, was investigated.

The third contribution of this thesis is the investigation of the efficacy of 3D TA as a means of predicting the survival prognosis of paediatric medulloblastoma: the most common malignant brain tumour occurring in children. To the best of my knowledge at the time of writing, there has been no published work on investigating

brain tumour survival predictors based on image analysis of conventional MRI, such as T1 and T2-weighted scans.

1.4 Thesis Structure

The chapters of this thesis are, to a large extent, self-contained and can be read independently. Chapters 2 to 4 present the context of the work and theoritical background that is of direct relevance; as well as the current state-of-the-art. Chapters 5 to 7 are experimental and offer the main contributions of this thesis. A summary of the thesis is discussed below:

Chapter 2: Background on MR Imaging of Paediatric Brain Tumours
This chapter gives a background on the neuroimaging of paediatric brain tumours. This begins with an introduction to the principles behind magnetic resonance imaging (MRI), followed by a discussion of important MR imaging parameters and how they are linked to brain tumour visualisation. This includes a review of the characteristics of the most common childhood brain tumours: medulloblastoma, pilocytic astrocytoma and ependymoma, and how they appear on conventional MR imaging.

Chapter 3: Background on Machine Learning
Since the diagnostic and prognostic classification of brain tumours from MR images is a classical machine learning problem, this chapter reviews relevant concepts from the field of machine learning. The emphasis of the chapter is on discussing supervised learning methods. In particular, it reviews common feature selection algorithms, classifiers and ways of validating the performance of classification models.

Chapter 4: TA of MR Images: Theory and State-of-the-Art

Here, an explanation of texture analysis (TA) methods, currently available in the literature, is presented, with a focus on statistical TA methods. Next, an extensive literature review of how TA was previously applied on MR imaging for the diagnostic classification of tumours is included, with a particular emphasis on brain tumours in children and adults. Whilst the MRI literature does not include any work on the application of TA for predicting survival prognosis in paediatric oncology, there have been recent efforts in other problem domains, such as computed tomography (CT) imaging of breast cancer. Such work is therefore reviewed at the end of the chapter for completeness.

Chapter 5: A Single Centre Study on 3D TA

This chapter is experimental and investigates the efficacy of 3D TA, within a supervised learning framework, to diagnostically classify the three most common types of brain tumours in children: medulloblastoma, pilocytic astrocytoma and ependymoma. As part of this study, a comparison of 3D TA to the traditional 2D approach is included. Additionally, the performance of six different classifiers was studied, in order to determine whether the choice of learning algorithm has a significant effect on diagnostic tumour classification performance. This study offers the first novel contribution of the thesis.

Chapter 6: A Multicentre Investigation on the Transferability of TA

This chapter is experimental and offers the second contribution of the thesis. The chapter presents a multicenter study on the efficacy and cross-center transferability of 3D TA as a diagnostic characterisation method, using MRI datasets obtained

from three different hospitals. This analysis also included a comparison of the performance of different feature selection methods as well as an investigation on the nature of features that are most likely to train classifiers that can generalise well with the 3D textural data.

Chapter 7: Predicting Survival in Paediatric Medulloblastomna

This chapter offers the third contribution of the thesis. Here, we look into the problem of predicting survival prognosis of brain tumours, and discuss an experiment that aims to establish the value of MRI TA as a potential prognostic marker. To this end, the study made use of clinical MRI datasets of patients diagnosed with medulloblastoma: the most common brain tumour occurring in children. This is clinically important, as long-term benefits of such prognostic marker could include personalised determination of treatment options and advancing clinical management of patients for therapy.

Chapter 8 presents a summary of achievements and overall conclusion, followed by suggestions for future work.

Appendix A: A Preliminary Study

During the early stages of this research, a preliminary classification experiment was conducted to explore the feasibility and efficacy of carrying out MRI TA in paediatric settings, using the conventional 2D approach. Due to the preliminary nature of this study, and the succinctness in its statistical analysis, it was not included in the main body of the thesis. The positive findings of the study, however, motiavted rigorious analysis of MRI TA for tumour characterisation.

Appendix B: Non-Statistical TA Techniques

Although the technical work presented in chapters 5-7 of this thesis is based on statistical TA techniques, two common non-statistical methods are introduced here as they had been used in a number of relevant studies in the literature. They were also used in the preliminary study presented in Appendix A.

Chapter 2

Background on MR Imaging of Paediatric Brain Tumours

2.1 Introduction

This chapter gives a background on the MR imaging of paediatric brain tumours. The first section introduces the principles on which MRI is based, using the classical physics description of the nuclear magnetic resonance (NMR) phenomenon. The material presented here is an outline of basic theory; for a more in-depth discussion, please refer to the excellent textbooks on the subject by ZP Liang and PC Leuterbur [1] and R Freeman [2]. The second section follows by providing an overview of the most common brain tumours in children, with a focus on their MR imaging characteristics.

2.2 Background on MR Imaging

2.2.1 Nuclear Magnetic Resonance

The principles on which magnetic resonance imaging (MRI) is based can be understood by appreciating the NMR phenomenon. In order to understand the mechanisms underlying NMR, it is necessary to start from the very centre of the atom: the nucleus. For a nucleus to generate an NMR signal, it must have nuclear spin: a fundamental property present in nuclei with odd atomic weights or odd atomic numbers. For the context of this thesis, our focus will be on Hydrogen atoms (1H nuclei), since over 70% of the human body is constituted of water. A spinning proton creates an electric current, and therefore a magnetic field, causing it to behave like a microscopic bar magnet (hence the use of the term magnetic in NMR). In other words, a spinning nucleus possesses angular momentum $\vec{J}$ and a charge, which give rise to an associated magnetic moment $\vec{\mu}$ [1]. Angular momen-

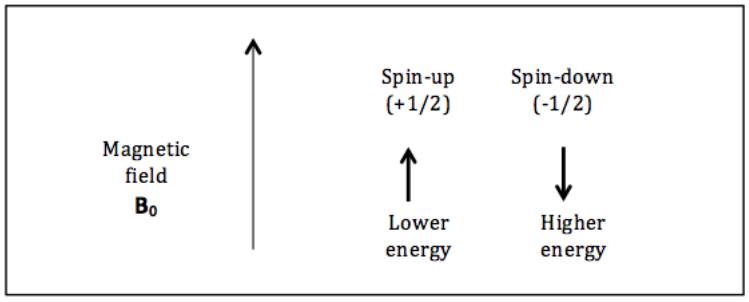


Figure 2.1: A diagram showing spin orientations when placed in an external magnetic field $\vec{B_0}$.

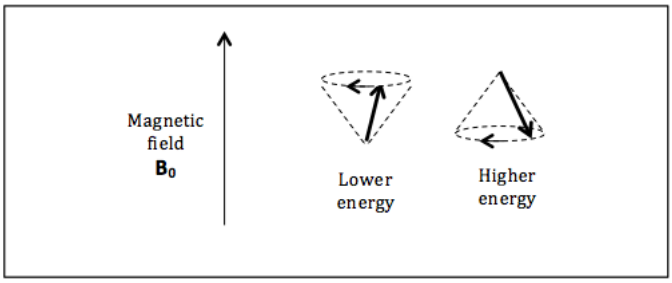


Figure 2.2: A diagram showing how spins precess around the axis of an applied magnetic field $\vec{B_0}$.

tum and magnetic moment are related to each other by:

$$\vec{\mu} = \gamma \vec{J} \tag{2.1}$$

where γ is a characteristic constant of the nucleus, known as the *gyromagnetic ratio*.

When placed in a magnetic field $\vec{B_0}$, the magnetic moment of a nucleus will orient relative to the field. The number of possible orientations is determined by the spin quantum number I. For a nucleus with spin number I, there are 2I+I possible spin states [2]. 1H has a spin number of $\frac{1}{2}$ and therefore has two possible spin states, denoted as $+\frac{1}{2}$ and $-\frac{1}{2}$. As shown in Figure 2.1, nuclei aligned parallel to the external field are in a lower energy state ($+\frac{1}{2}$ 'spin-up'), whereas those aligned anti-parallel to the external field are in a higher energy state ($-\frac{1}{2}$ 'spin-down').

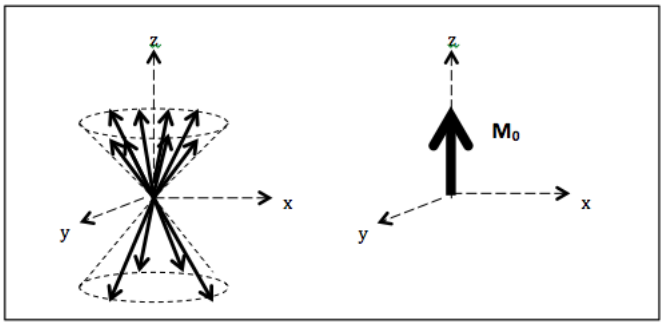


Figure 2.3: A vector model of NMR representing spin orientations and magnetisation vector $\vec{M}_0$.

1H nuclei precess around the axis of the applied static magnetic field $\vec{B}_0$, as depicted in Figure 2.2. The frequency (ω), at which the nucleus precesses is determined by the strength of the applied magnetic field B_0 and the nucleus' gyromagnetic ratio γ. The frequency of precession is called the *Larmor frequency* and is mathematically described in Equation 2.2:

$$\omega = \gamma \vec{B}_0 \tag{2.2}$$

In NMR experiments, we are not interested in the behaviour of individual spins, but rather in the collective behaviour of the overall spin system. There is an excess of a very small fraction of spin-up nuclei; this is because a spin is more likely to take the lower-energy state (with higher stability) than the higher energy state. By summing over the spin orientations, an overall macroscopic magnetisation vector $\vec{M}_0$ is produced and is aligned along the direction of the applied magnetic field (Equation 2.3), as depicted in Figure 2.3. It is the manipulation of $\vec{M}_0$ that forms the basis of all MRI scans.

$$\vec{M} = \sum_{i=1}^{N_s} \vec{\mu}_i \tag{2.3}$$

Where N_s is the total number of spins.

2.2.2 Radio Frequency Pulses

Since the magnetisation vector $\vec{M}_0$ is very small (on the order of microtesla), it is difficult to measure at its equilibrium state while aligned with the applied magnetic field $\vec{B}_0$. The magnetisation vector needs to precess orthogonally to the applied magnetic field; the generated current can then be detected by the receiver coil, which can be placed at right angles to the axis of the applied magnetic field. In order to achieve this, radio frequency (RF) pulses are applied at the Larmor frequency.

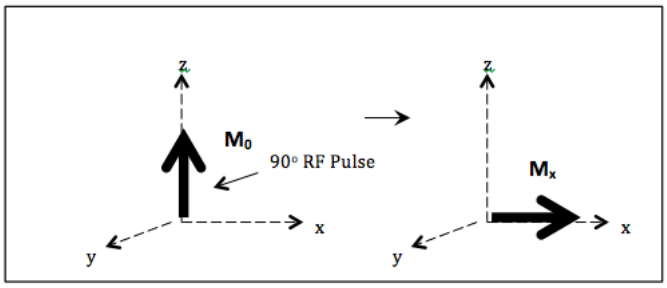


Figure 2.4: A diagram showing a 90 degree RF pulse being used to manipulate $\vec{M}_0$ into the XY plane.

The intensity and duration of the applied RF pulse are chosen so that $\vec{M}_0$ rotates by 90 degrees, being flipped from its equilibrium state at the z-axis to the XY plane. In this position, the magnetisation vector can induce maximum signal in the receiver coil. Figure 2.4 shows the manipulation of $\vec{M}_0$ into the XY plane. The degree by which $\vec{M}_0$ is flipped (θ) can be calculated as shown in Equation 2.4:

$$\theta = \gamma \vec{B}_1 t_p \tag{2.4}$$

Where $\vec{B}_1$ is the generated RF field and t_p is the time period during which the pulse is applied.

2.2.3 Rotating Frame of Reference

In NMR, a *rotating frame of reference* is used to simplify the complex motion of precessing spins. In this frame of reference, the XY plane is assumed to rotate around the z-axis, at the Larmor frequency, causing the magnetisation vector and $\vec{B_1}$ to appear stationary.

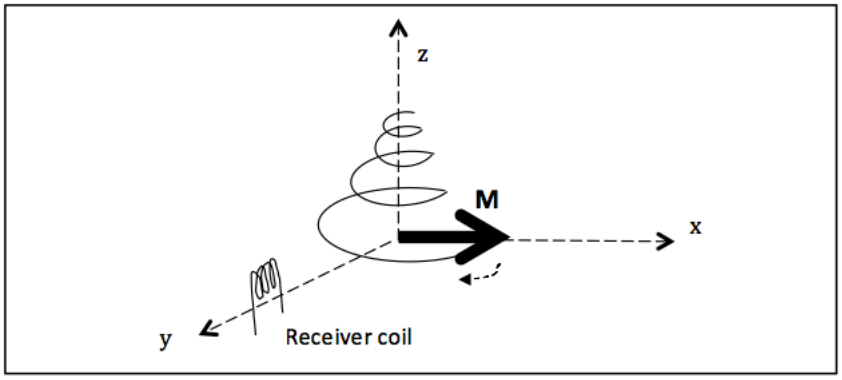


Figure 2.5: A diagram depicting how, after the 90 degree excitation pulse, the magnetisation vector gradually returns to its equilibrium state.

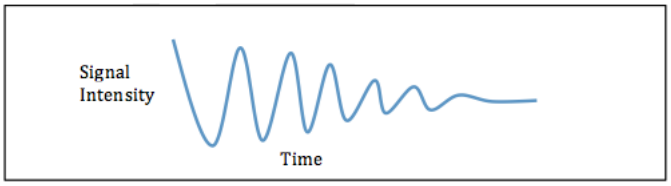


Figure 2.6: A plot showing how the free induction decay (FID) signal decays with time as the system returns to equilibrium.

While the magnetisation vector is in the XY plane, it rotates and consequently induces a weak, oscillating voltage in the receiver coil. This observed voltage corresponds to the detected MR signal and is proportional to the transverse magnetisation $\vec{M_{XY}}$, as characterised by the following equation:

$$V(t) \propto \int \frac{\delta \vec{M_{XY}}(t)}{\delta t} \delta r \tag{2.5}$$

Where δr is the volume element given by $\delta x \delta y \delta z$.

The signal, after processing, can be given in the following generalised form:

$$S(t) = \int \vec{M_{X'Y'}}(t)\delta t \tag{2.6}$$

This magnetisation does not remain in the XY plane, however, but gradually returns to its original equilibrium state (Figure 2.5). The detected NMR response is referred to as the *free induction decay (FID)*, which is depicted in Figure 2.6. FID is caused by two distinct mechanisms, namely longitudinal and transverse relaxation. The two processes are explained in the next section.

2.2.4 Longitudinal (T1) Relaxation

Longitudinal relaxation is the recovery of the magnetisation vector in the z-axis ($\vec{M}_z$) back to its equilibrium value ($\vec{M}_0$) following perturbation, and is caused by the loss of energy from spins to the surrounding environment. T1 is the time constant used to represent this process and can be defined as the time taken for a signal to recover back to 63% of its original value. This process is exponential and can be mathematically described as shown in Equation 2.7:

$$\vec{M}_z = \vec{M}_0(1 - \exp(-t/T1)) \tag{2.7}$$

where t is a time delay that is used to allow some longitudinal relaxation to occur, as explained in the next paragraph.

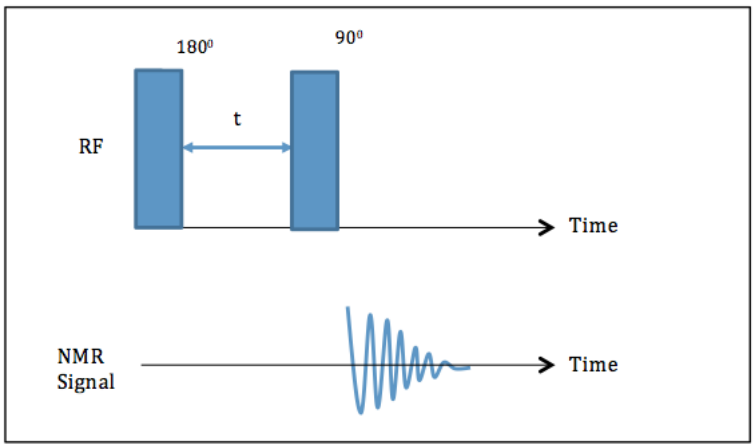


Figure 2.7: A diagram showing the pulse sequence timings for an inversion recovery experiment to measure longitudinal relaxation (T1).

T1 can be measured using an inversion recovery pulse sequence. In inversion recovery, a 180 degree pulse is initially applied, causing the magnetisation vector to be inverted onto the negative z-axis. A time delay, t, is then introduced, during which some magnetisation is returned back to the positive z-axis. After t, a 90 degree pulse is applied, leading to the rotation of the magnetisation vector onto

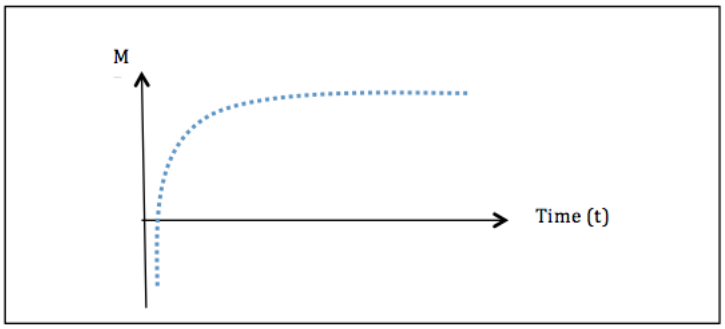


Figure 2.8: A plot depicting the relationship between the recovery of longitudinal magnetisation and time delay in an inversion recovery experiment.

the XY plane. At this point, the signal amplitude of the FID can be recorded, as depicted in Figure 2.7. Finding T1 involves measurements of the signal amplitude for several different t intervals, as depicted in Figure 2.8. Performing the curve fitting according to Equation 2.7 will give T1.

2.2.5 Transverse (T2) Relaxation

T2 is the relaxation of spins in the transverse (XY) plane. Upon applying a 90 degree pulse, the spins initially align in the XY plane and have phase coherence. However, for this phase coherence to be maintained, all spins need to experience the same magnetic field. The presence of static magnetic field heterogeneities and interactions with neighbouring molecules lead to loss in phase coherence. This process is exponential in behaviour, as shown in Equation 2.8, and is always shorter than T1 .

$$\vec{M_{XY}} = \vec{M_0} \exp(-t/T2) \tag{2.8}$$

One way T2 can be measured is through the use of *Spin Echo (SE)* pulse sequence [3]. Here, a 90 degree RF pulse is first applied, bringing the magnetisation vector into the XY plane. While in the XY plane, the spins start to de-phase. A

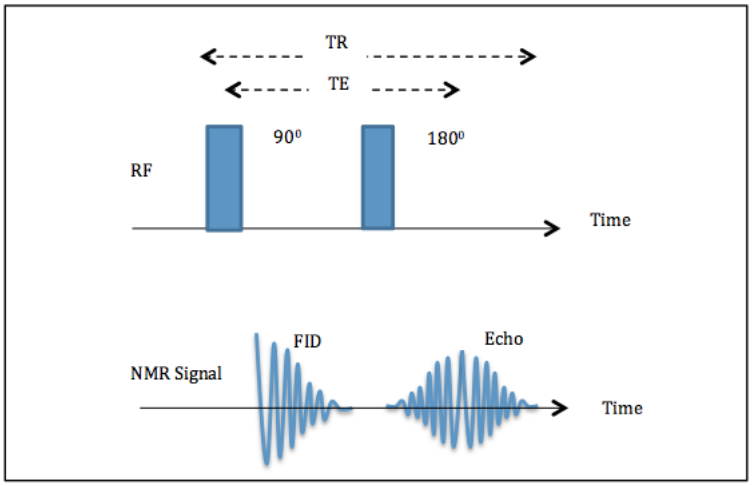


Figure 2.9: A diagram showing the pulse sequence timings for a spin echo (SE) experiment. The time to echo formation is referred to as echo time (TE), while the time between successive excitations is referred to as repetition time (TR).

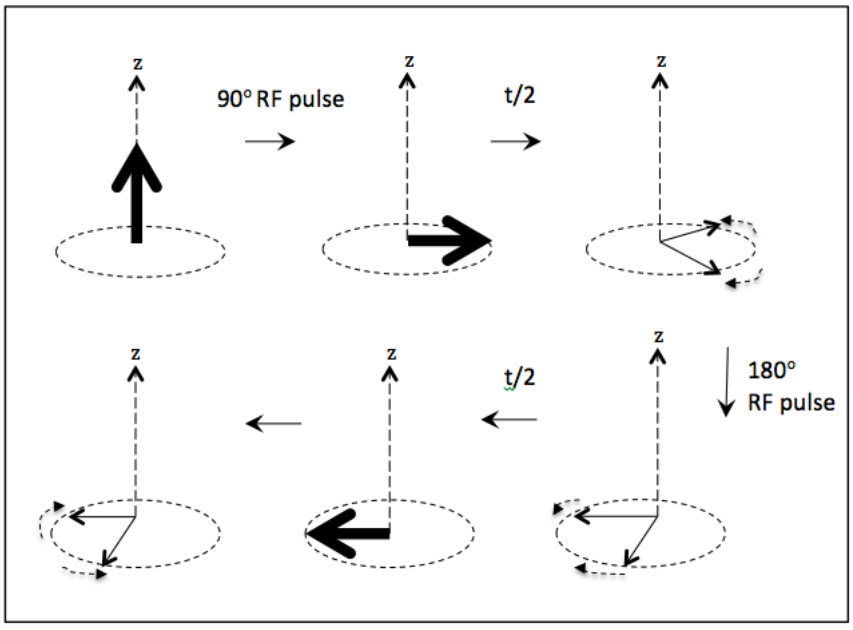


Figure 2.10: A diagram showing the de-phasing and refocusing of the magnetisation vector during a spin echo pulse sequence.

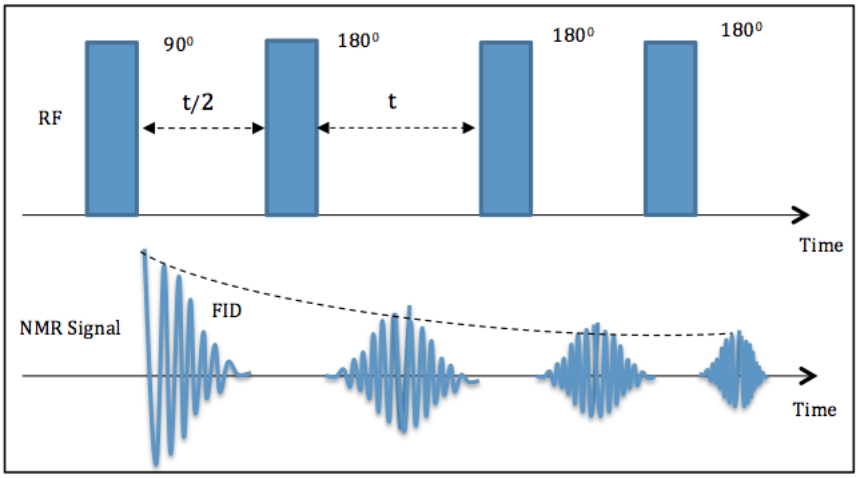


Figure 2.11: A diagram showing the formation of spin echoes by multiple 180 degree pulses (the pulse sequence and the resulting echo signals). .

180 degree pulse is then applied, causing the de-phasing spins to rotate around the XY plane. Although they continue to de-phase, the direction in which these spins do so now acts to refocus the magnetisation, thereby forming an 'echo'. The time to echo formation of the spin-echo signal is referred to as *echo time (TE)*, while time between successive excitations is referred to as *repetition time (TR)*. The formation of an echo in a SE experiment is depicted in Figures 2.9 and 2.10.

When a spin system is excited by a 90 degree pulse followed by a sequence of 180 degree pulses, a train of spin echoes will be generated. Suppose that the 90 degree pulse is applied at t=0 and the 180 degree pulses at (2n-1)t for n=1,2,..N. A train of N echoes will be formed at time 2nt. The echo amplitudes are characterised by Equation 2.9, below:

$$E_n = exp(-2nt/T_2) \tag{2.9}$$

Because of the straightforward relationship for the echo amplitudes, as per Equation 2.9, this multiple-echo sequence is an efficient way to measure T2 values. This sequence is known as the *CPMG (Carr-Purcell-Meiboom-Gill)* sequence and

is widely used in practice (Figure 2.11).

2.2.6 Magnetic Field Gradients

So far, it has been explained how an NMR signal can be induced through the application of a static $\vec{B_0}$ and an RF pulse, in order to flip the magnetisation into the XY-plane. However, for the signal to be useful for large inhomogeneous samples (such as the human brain), it has to be manipulated so that a signal from a particular region of the sample has properties that distinguish it. To illustrate how this is achieved in MR, the concept of *magnetic field gradients* needs to be introduced.

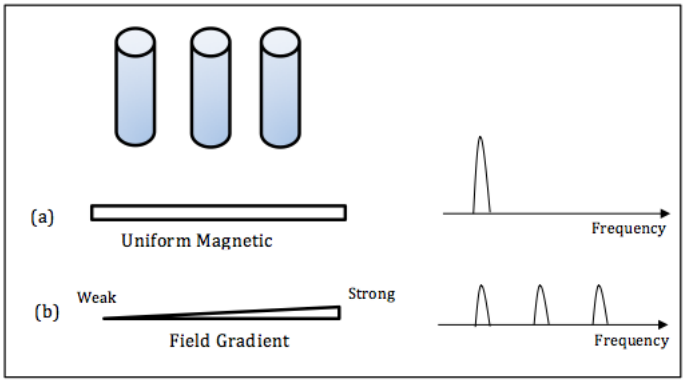


Figure 2.12: A diagram showing three test tubes filled with water and how the use of an external magnetic field affects the resonant frequencies across the tubes (a) The use of a uniform magnetic field yields only one frequency (b) the use of a magnetic field gradient causes the resonant frequencies to be different across the tubes. Signals obtained from different test tubes have therefore been spatially localised by the field gradient.

Consider an example where we have three test tubes filled with water, as shown in Figure 2.12. Now introduce three different magnetic field strengths and arrange the test tubes so that the strength of the applied magnetic field increases from one sample to the next (that is, apply a field *gradient*). Because of the variations on the strengths of the applied magnetic field, the Larmor frequencies across the

test tubes become different. In other words, the NMR signal was made spatially dependent.

Similarly, in MR imaging, spins at different spatial locations would be excited in the same way if they resonate at the same frequency. However, if each of the regions of spins was to experience a unique magnetic field, it would be possible to spatially localise them and image their positions. Hence, the key to producing an MR image is the fact that the NMR frequency is strictly proportional to the strength of the magnetic field experienced by the spins. Gradients are produced using coils through which an electric current is passed to induce a particular local magnetic field. By applying a magnetic field gradient, the Larmor frequency at position r becomes dependant on $\vec{G}$, the pulsed field gradient component parallel to $\vec{B_0}$, as shown in Equation 2.10:

$$\omega(r) = \gamma\vec{B}_0 + \gamma\vec{G} \cdot r \tag{2.10}$$

In MRI, the concept of field gradients is exploited in three successive steps: slice selection, frequency encoding and phase encoding, which are explained below.

2.2.7 Slice Selection

As discussed before, magnetic field gradients are used as a means of encoding spatial information in MRI. Consider an example where we apply a field gradient along the main axis of the human body (in this discussion, this axis is referred to as the z-direction). The gradient causes the spins to have different frequencies across the axis of the body, thus "dividing" the body into a number of "parallel slices", each having a unique Larmor frequency. A slice of interest can then be chosen for further examination, as detailed in the next paragraph.

An MRI slice needs to have a certain thickness that is enough to provide a suf-

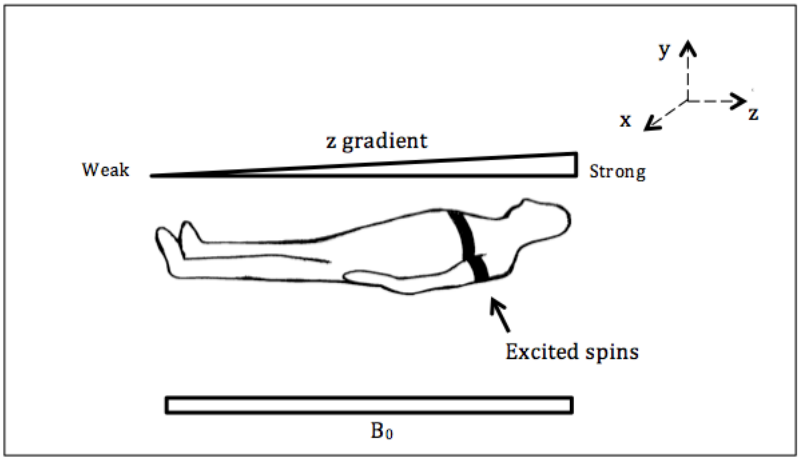


Figure 2.13: A diagram showing slice selection being carried out on a patient using a magnetic field gradient and an RF pulse.

ficient MRI signal, but not so thick that the tissue structure changes significantly across the limited dimension [2]. Slice selection involves the use of an RF pulse that has low intensity and long duration. We can select just one slice by adjusting the frequency-band of the selective RF pulse, so that only these spins are excited. Figure 2.13 illustrates the use of a field gradient and an RF pulse for selecting a slice from the patient's body. Outside the chosen slice, spins are far enough from the transmitter frequency that any excitation they experience becomes negligible.

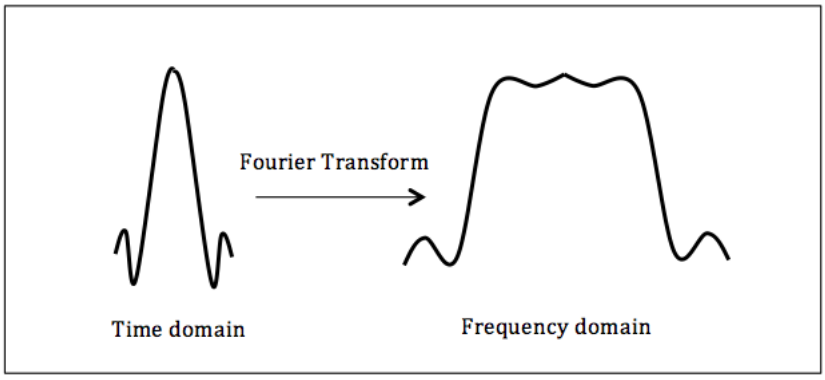


Figure 2.14: A diagram showing the time and frequency excitation profiles of the desired RF excitation pulse..

Ideally, the RF pulse needs to be a Sinc function (Equation 2.11) in the time

domain and a rectangular pattern in the frequency domain.

$$Sinc(x) = Sin(x)/x \tag{2.11}$$

The use of a rectangular pattern with steep edges in the frequency domain makes it straightforward to choose a well-defined frequency-band to excite spins in the desired slice. It is worth noting that in practice, the achieved profile of the RF pulse is a rough approximation of the ideal pulse (Figure 2.14). Nevertheless, the edges are reasonably steep, and there is only weak excitation of signals from adjacent regions.

The thickness of the slice is determined by the effective bandwidth of the RF pulse and the strength of the applied gradient. The relationship between slice thickness (δz), RF frequency bandwidth (δF) and the slice-select gradient $\vec{G}$ is characterised by Equation 2.12. Stronger gradients produce thinner slices, and vice versa.

$$\delta F = \gamma \cdot \vec{G} \cdot \delta z \tag{2.12}$$

2.2.8 Frequency and Phase Encoding

Suppose that slice selection has been carried out by applying a field gradient along the z-axis; the next step is to determine the spin density within the selected two-dimensional sample with respect to the two remaining axes (referred to here as x and y directions). This is done via *frequency and phase encoding*. In frequency encoding (also known as "readout"), a gradient is switched on during acquisition and the signal is acquired at fixed time intervals. As long as the gradient is on, the nuclei experiencing different field strengths will precess at different frequencies. Hence, their location in the x-direction can be encoded in terms of the frequency

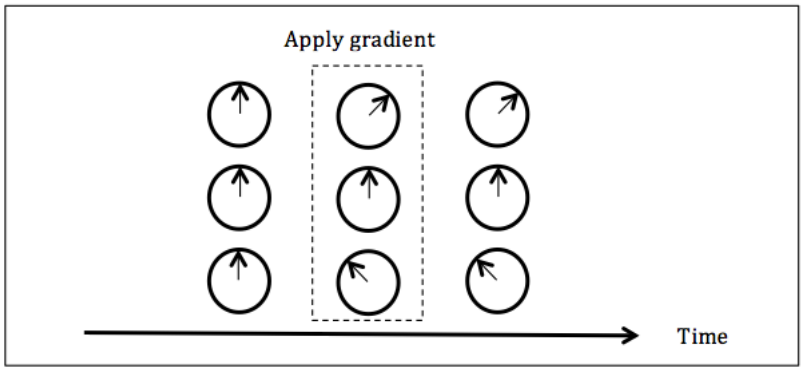


Figure 2.15: A diagram illustrating the use of a field gradient for carrying out phase encoding. Prior to applying a gradient, the spins precess with the same phase. When a gradient is applied, the spins will have different energies and hence get out of phase. Removal of the gradient brings the spinning frequency back to the original value, however, the phase different is reserved.

of the acquired signal.

In phase encoding, a magnetic field gradient is applied in the y-direction during a fixed time period, but is switched off prior to signal acquisition. Switching off the gradient means that the frequencies of all spins become very similar; however, their phases remain different. The use of a field gradient to carry out phase encoding is depicted in Figure 2.15. By combining frequency and phase encoding, we are able to acquire the signal for all voxels in the selected slice. The obtained signal is comprised of a number of sinusoids at different frequencies and phases, each representing signal obtained from a unique location. Thus, Fourier Transform can be used to convert the signal in terms of the unique sinusoids it comprises.

2.2.9 K-Space

As mentioned earlier, an imaging sequence can be made up of three stages: 1. Slice selection in a constant z-gradient. 2. Frequency encoding (readout) in a constant x-gradient. 3. Phase encoding in a stepped y-gradient.

The collected raw data is stored in a matrix known as k-space. The original formulation of k-space is given by the following equation [62].

$$k = \int_0^t \gamma \vec{G}(t)\delta t \tag{2.13}$$

where $\vec{G}(t)$ is a time-dependent gradient applied after slice selection.

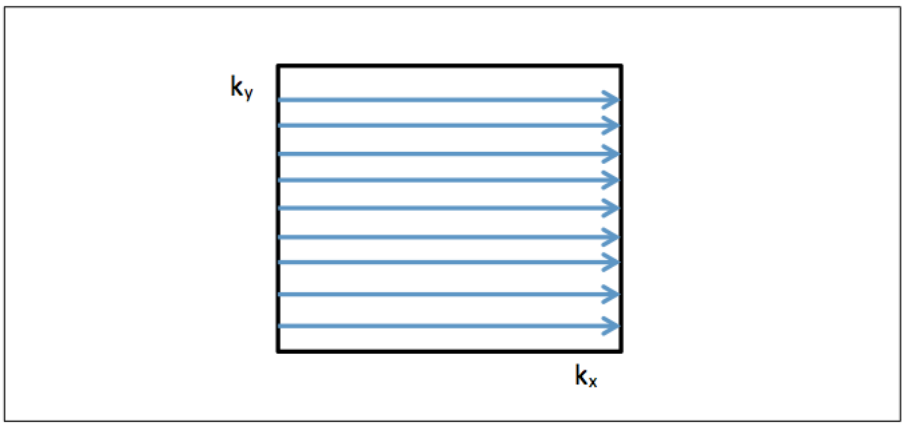


Figure 2.16: A diagram showing how the overall k-space trajectory may be a set of horizontal traces stacked one above the other.

Whilst three dimensions are involved, we simplify the discussion by focusing on on the k_x and k_y axes of k-space, with the assumption that slice selection was carried out. The way data gathering is organised to explore k-space is called *trajectory*. In the simplest case, the readout information is acquired as a function of real time along the k_x axis. The k_y information, however, is acquired through a series of separate measurements. This involves varying the phase-encoding strength, causing each line to assume different k_y locations, which results in a rectilinear

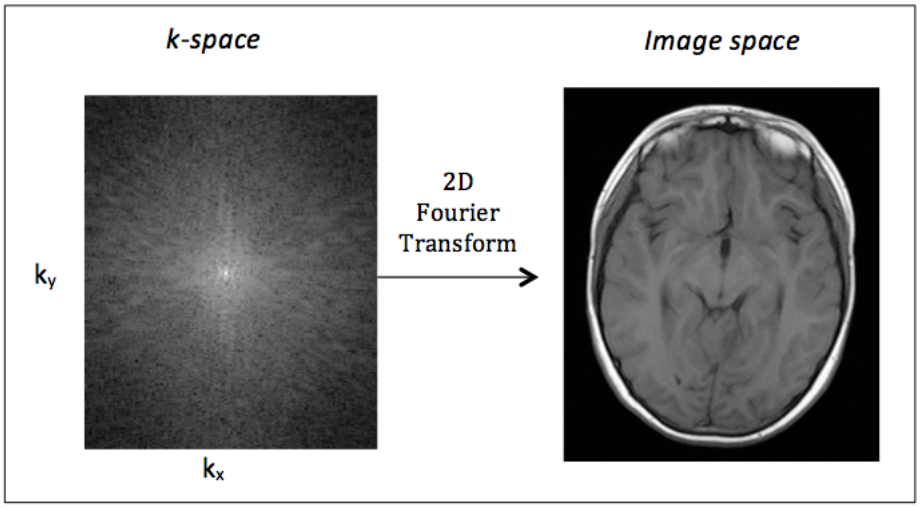


Figure 2.17: A diagram showing how 2D Fourier Transform can change MR data of a human brain in k-space to image space. Original image space data was obtained from the CCLG database [4]

sampling as shown in Fig 2.16.

K-space data can finally be converted to *image space* via two-dimensional Fourier transform. An example of data in k and image space is shown in Figure 2.17. Note that each individual point in k-space represents one measurement of the entire NMR signal. Hence, each point in k-space contributes information to all the pixels in image space.

2.2.10 Imaging Planes

Using the aforementioned practices of slice selection, frequency encoding and phase encoding, MRI can create images in different anatomical planes, enabling study of structures from different viewpoints. The three primary imaging planes that are utilised in MR imaging are *axial, sagittal and coronal*. Axial sections form a series of slices that run top (superior) to bottom (inferior). Sagittal sections run from one side of the body to the other: left to right or right to left. Coronal slices follow front (anterior) to back (posterior), as though cutting through a halo around the

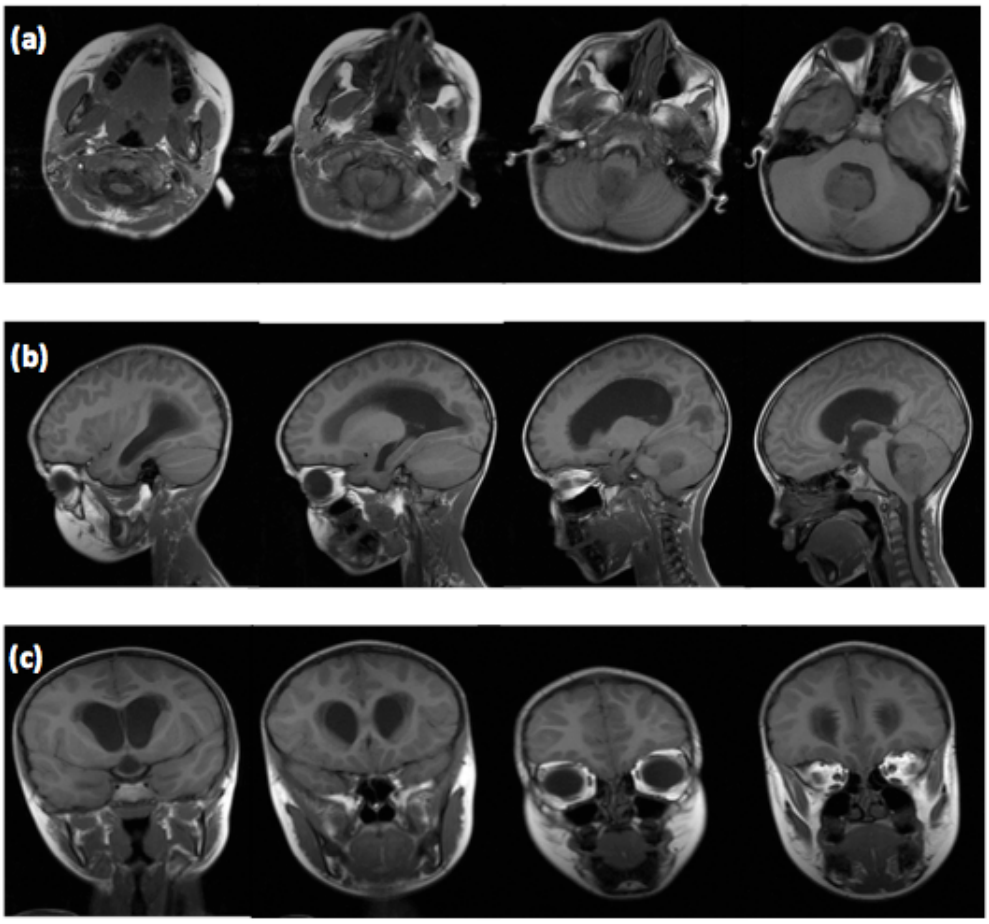


Figure 2.18: A diagram showing a number of brain MR image slices in three different planes: (a) axial, (b) sagittal and (c) coronal. Original images were obtained from CCLG database [4]

structure being visualised. Figure 2.18 shows a number of brain MR images that belong to the same subject in the three different planes.

2.2.11 Image Contrast

Since an important application of MR imaging is to distinguish between regions of different compositions (e.g. different biological tissues), the concept of *contrast* needs to be introduced. Contrast is produced when there is variation in signal intensity between image pixels and can be created in an image using a number of ways, including: 1. Spin density. 2. T1 or T2 relaxation times. In a sample, regions with large numbers of excited spins have high signal intensities. For MRI, images produced using spin density contrast can be referred to as *proton density maps*; such images appear brighter in regions with higher spin density and darker in regions with lower spin density. However, the contrast of the image is not fixed and can be manipulated in order to favour important clinical features.

In order to illustrate how image contrast can be manipulated using T1 and T2 relaxation time parameters, consider the example of cerebrospinal fluid (CSF), which has T1 relaxation time that is longer than average: between 2 and 4 seconds. The average white matter T1 relaxation time is about 0.4 seconds; if consecutive measurements are carried out with a repetition time (TR) = 0.4, a steady-state system will be established[1], with CSF 1H spins not completely recovering and thereby producing weak CSF signal. CSF will therefore appear dark on *T1- weighted*. Regions that appear brighter on T1- weighted images (e.g. fat) have relatively short T1 relaxation times.

For an image to be *T2-weighted*, the TR time used would need to be significantly long to remove T1-weighting; and the echo time (TE) used would also need to be long to introduce T2 contrast. Regions with bright intensities on T2-weighted images represent high relaxation times, which means that signal arising from CSF would appear brighter on T2 images. Example T1 and T2-weighted

[1]If a system is in steady state, then the recently observed behaviour of the system will continue into the future

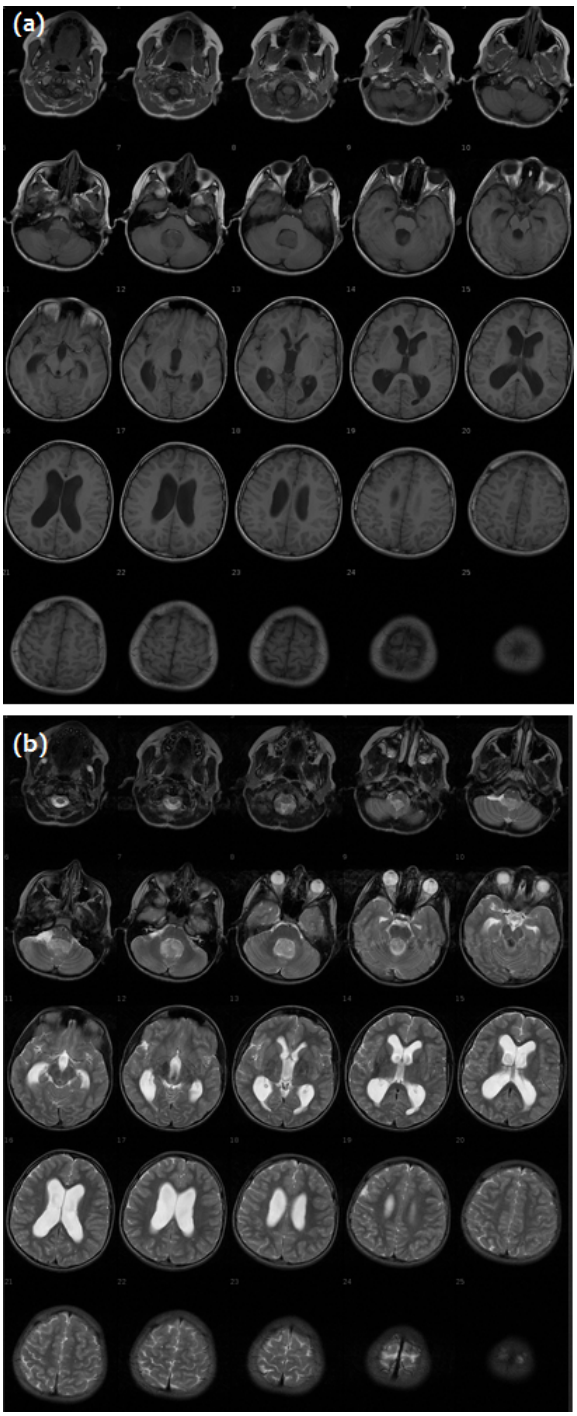


Figure 2.19: (a) T1 and (b) T2-weighted brain MR images demonstrating how image contrast can be manipulated using T1 and T2 relaxation times. TR=414 ms and TE=17ms for T1, whereas TR=6980ms and TE=77ms for T2. A 1.5 T scanner was used to acquire the images. Original images were obtained from CCLG database [4].

brain MR images are shown in Figure 2.19.

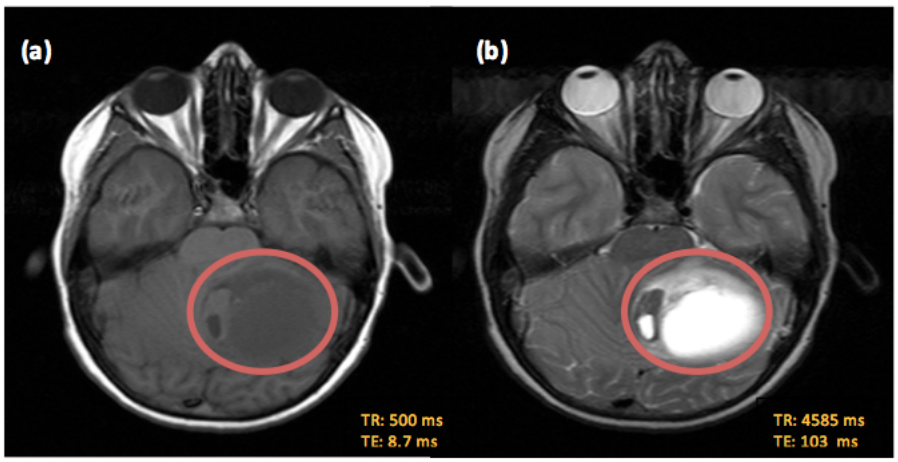


Figure 2.20: (a) T1 and (b) T2-weighted brain MR images demonstrating how image contrast can be manipulated using T1 and T2 relaxation times. A 1.5 T scanner was used to acquire the images. Original images were obtained from CCLG database [4].

Figure 2.20 shows T1 and T2-weighted MR images of a child diagnosed with brain cancer. For the T1-weighted image, the TR and TE values used were 500ms and 8.7ms respectively, whereas for T2-weighted image, the TR and TE values used were 4585ms and 103ms respectively. One can see how different tumour regions (marked in red in Figure 2.20) can be visualised with different contrasts characteristics on both images. For instance, the large cystic mass of the tumour appears brighter than the solid region on the T2-weighted image, whereas the opposite is true for the T1-weighted image. This illustrates how varying scanning parameters can manipulate the contrast of different brain tumour components, which can aid the visualisation for clinical decision-making. The next section of this chapter elaborates on how common childhood brain tumours appear on conventional MR imaging.

Note that in some occasions it can be advantageous to enhance image contrast by injecting the patient with *contrast agents*. Examples of such chemicals include

complexes that contain paramagnetic ions, which accelerate T1 relaxation by providing local magnetic fields that fluctuate near the Larmor frequency. The most commonly used agents are those containing gadolinium Gd+3 ions, but the uptake of contrast agents is generally specific to a particular tissue type or pathology.

2.3 MR Imaging Characteristics of Brain Tumours in Children

This section provides an overview of the most common brain tumours in children, with a focus on their MR imaging characteristics. It starts by introducing basic brain anatomy and then moves on to present commonly occurring brain tumours and how they appear on T1 and T2-weighted images.

2.3.1 Basic Brain Anatomy

The main parts of the brain are:

- Cerebrum: This forms the largest part of the brain and is located at the top. It comprises two hemispheres and controls higher functions: thinking, learning, problem solving and emotions.
- Cerebellum: This is the back of the brain and controls balance, movement and coordination.
- Brainstem: This controls automatic functions, i.e. breathing, body temperature, heart rate, blood pressure, eye movement and swallowing. It is located in the lower part of the brain and provides a connection to the spinal cord [5].

The cerebrum has a folded surface called the cortex, which contains about 70% of the 100 billion nerve cells. The cortex contains neuron cell bodies (*grey-matter*), which are interconnected to other brain regions by axons (*white-matter*).

The brain has fluid-filled cavities called ventricles. These contain a ribbon-like structure, choroid plexus, which produces the *cerebrospinal fluid (CSF)*. CSF flows within the brain and spinal cord to help cushion it from injury.

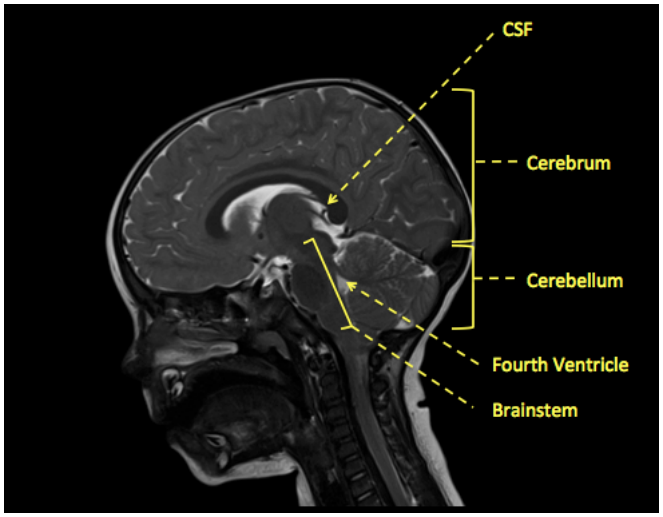


Figure 2.21: MR scan of a child showing brain structure. Original image was obtained from Siemens Healthcare webpage on Paediatric MRI [82] .

The area near the base of the skull is divided into three regions: posterior, middle and anterior fossae. *Posterior fossa* is the largest and contains the cerebellum and the brainstem. Tumours that arise here are of special interest because the posterior fossa is near critical brain structures, making any tumours present difficult to treat. Around 55% of childhood brain tumours arise in the posterior fossa, compared with 15% to 20% of adult tumours. Tumours occurring in childhood are more likely to be *primary* rather than secondary tumours, meaning that they originate from brain tissues rather than metastatic tumours that spread from a different body part [6].

2.3.2 Paediatric Brain Tumours

As previously discussed in Chapter 1, brain tumours are the second most common cancer in children. Brain tumours can be classified to over 30 classes, as per the WHO classification system [75]. The three most common posterior fossa tumours are *medulloblastoma* (MB), *pilocytic astrocytoma* (PA) and *ependymoma* (EP); the three types that are of particular interest within the context of this thesis. It

is worth noting that these tumour classes are grouped into broad histopathological categories. Under this classification, MBs are classed as *embryonal*, PAs are classed as *astrocytic* and EPs are classed as *ependymal* tumours. In this section, an overview of MB, PA and EP is given, together with a summary of how they manifest on conventional MR images.

Medulloblastoma

MB is the most commonly occurring posterior fossa tumour in children, accounting for up to 40% of the cases [7]. The peak ages for MB presentation are 3 and 7 years of age [8]. They are highly malignant tumours (WHO Grade IV) and are twice as likely to occur in male as female children [7]. There exists different pathological subtypes of MB: *classic, demoplastic, nodular, large cell and anaplastic*, with the classic subtype being the most common [9].

On T1-weighted scans, MBs are *isointense*[2] to *hyperintese* [3] compared to white-matter. On T2-weighted scans, however, their appearance is variable, depending on tumour cellularity. Tumour components that are more cellular appear *hypointense*[4], whereas less cellular components appear iso- (or mildly hyper-) intense [9]. Although MB usually grow in circumferential patterns and maintain round borders, more aggressive forms may penetrate regions, such as the fourth ventricle or the brainstem [10]. Figure 2.22 shows T1 and T2-weighted MR images of a child diagnosed with medulloblastoma.

In terms of prognosis, MB survival depends on a number of factors that include age at the time of presentation. Presence of CSF dissemination and presence of any residual tumour after surgery are also prognostic factors. The best prognosis

[2]Of a similar intensity to a reference structure.
[3]Of a higher intensity to a reference structure.
[4]Of a lower intensity to a reference structure.

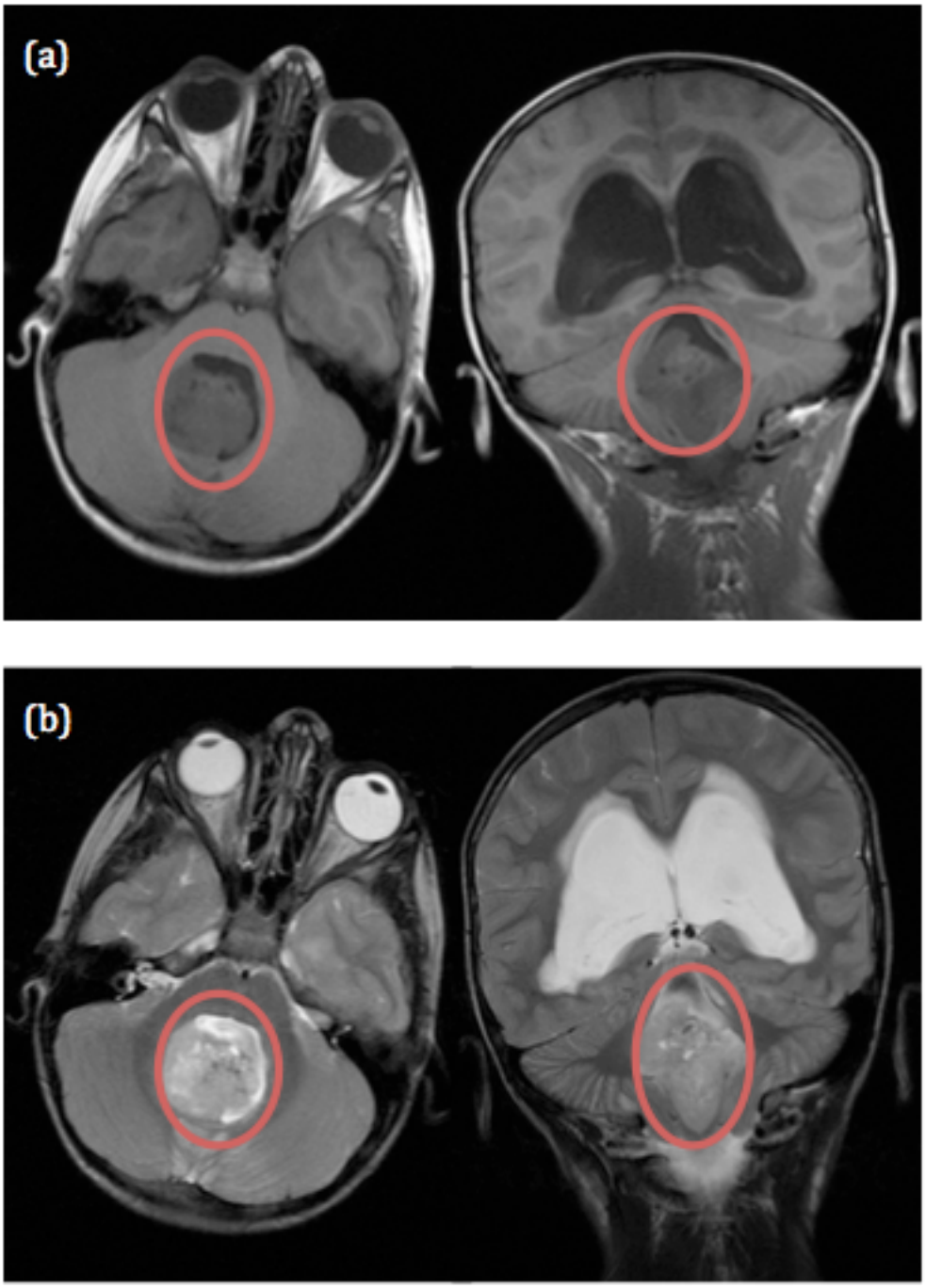


Figure 2.22: (a) T1- and (b) T2-weighted MR images of a child diagnosed with medulloblastoma. Images are shown in the axial and coronal planes. The red regions of interest indicate areas affected by the tumour. Original images were obtained from CCLG database [4] .

is about 80% for 5-year survival [5] and is found to be the case with children who are older than 3 years of age at the time of presentation [10], [11].

Pilocytic Astrocytoma

PA can affect different brain regions including the cerebellum, optic pathway and hypothalamus. Cerebellar PA account for one third of posterior fossa tumours in children, second to MB. The peak age of PA presentation is between 5 and 16 years of age. They are considered low-grade tumours (WHO Grade I); and boys and girls are equally likely to be affected by PA [10].

PAs are usually well confined and have a large cystic component [6] with a mural nodule [7]. On T1-weighted scans, both solid and cystic components appear as hypointense (similar to CSF signal intensity), while on T2-weighted scans, both components appear as hyperintense. Upon administering a contrast agent, the solid nodular component gets enhanced and enhancement of the cyst wall may occasionally be seen too. Figure 2.23 shows T1 and T2-weighted MR images of a child diagnosed with pilocytic astrocytoma. PA has an excellent prognosis of over 90% for 25-year survival [12]. In general, complete resection through surgery is considered curative [10] [11].

[5] 5-year survival rate is used for estimating the prognosis of a particular disease by describing the percentage of patients that are alive 5 years after their disease was diagnosed.

[6] A cyst is an abnormal sac that may contain air, fluids or semi-solid material.

[7] A mural nodule is a small lump of solid tissue on a cysts inner wall.

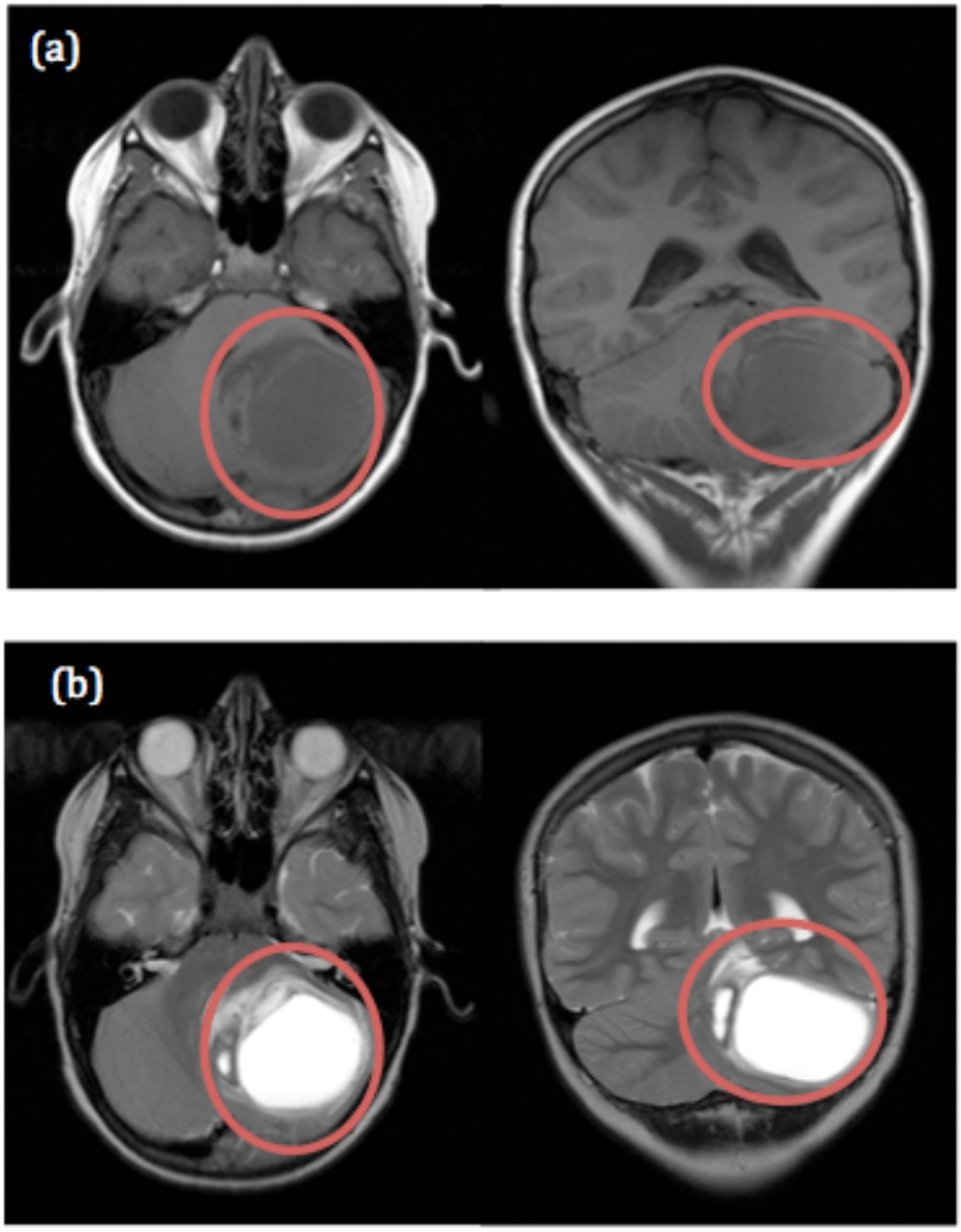


Figure 2.23: (a) T1- and (b) T2-weighted MR images of a child diagnosed with pilocytic astrocytoma. Images are shown in the axial and coronal planes. The red regions of interest indicate areas affected by the tumour. Images were obtained from CCLG database [4].

Ependymoma

Following MB and PA, EP is third most common childhood posterior fossa tumour, which usually occurs along the floor or roof of the fourth ventricle [10]. Most EPs are WHO Grade II with an average time of presentation at 6 years of age [13].

EPs appear iso- to hypointense on T1-weighted MRI and hyperintense on T2-weighted MRI. The solid components of EP demonstrate more heterogeneous signal characteristics compared to MB. After treatment, EP recurrence tends to be common due to tumour adherence to adjacent structures, which makes complete resection difficult [10]. Figure 2.24 shows T1 and T2-weighted MR images of a child diagnosed with pilocytic astrocytoma.

With regards to prognosis, the survival of posterior fossa EP is generally dependant on age, degree of initial surgical resection, presence of any dissemination and recurrence. EP′s survival rate is between 50% and 70% for 5-year survival [10] [11].

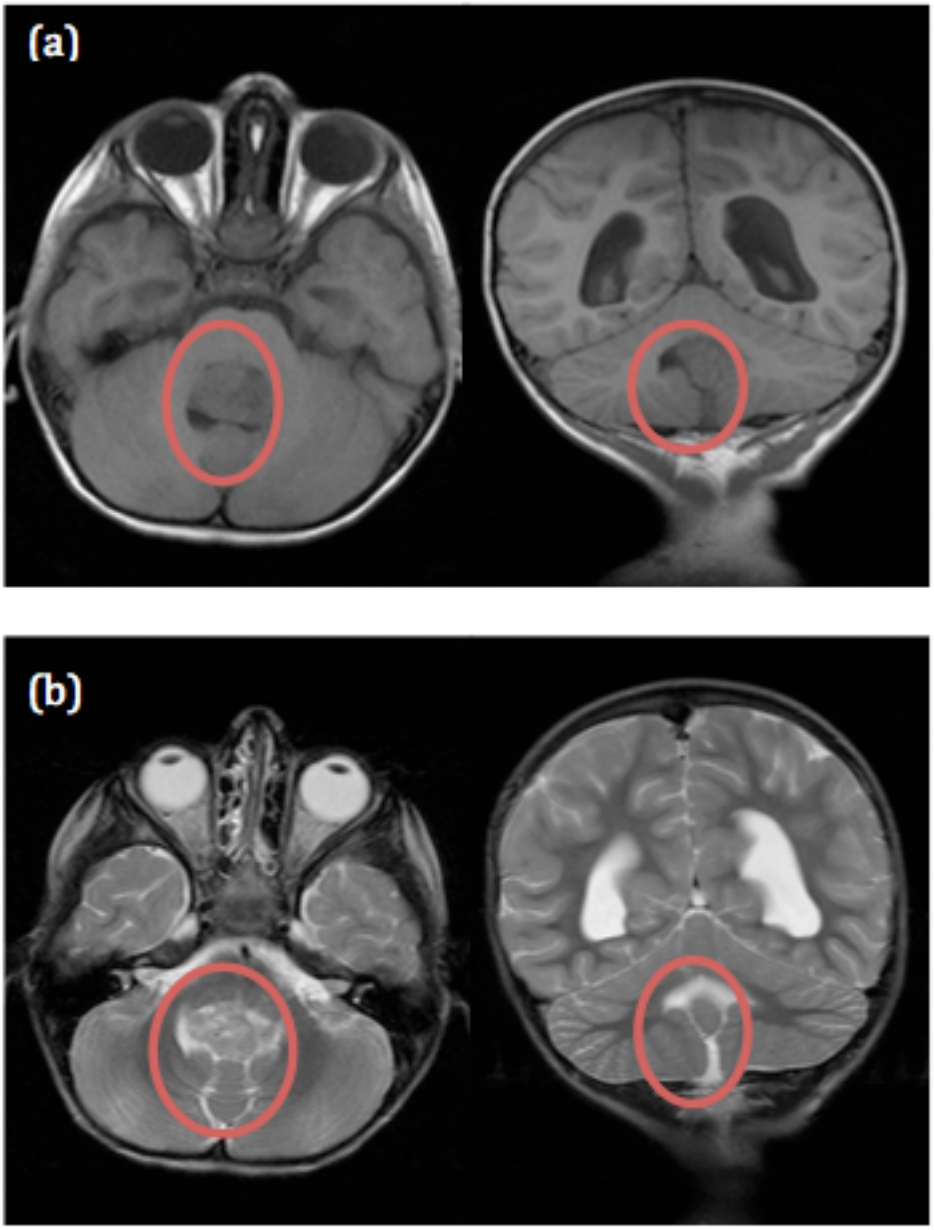


Figure 2.24: (a) T1- and (b) T2-weighted MR images of a child diagnosed with ependymoma. Images are shown in the axial and coronal planes. The red regions of interest indicate areas affected by the tumour. Images were obtained from CCLG database [4] .

2.3.3 Comparison to Adult Brain Tumours

It is worth noting that a number of differences exist between paediatric and adult brain tumours, in terms of their biology, occurrence and sensitivity to treatment. For instance, in childhood, cancer is related to tissue development and may initiate during the development of the embryo, as in the case of embryonal tumours like MB [91]. However, in adults, cancer is linked to the interaction of cells with environmental carcinogens. In addition to this, commonly occuring paediatric brain tumours include MB and PA, while in adults, glioblastoma and meningioma are the most common primary brain tumours [92]. Moreover, most paediatric brain tumours are primary, while those occurring in adults are more likely to have metastasised from other parts of the body as a result of other types of cancer, such as lung, breast and kidney cancer. In terms of treatment, children are more sensitive to radiotherapy and chemotherapy, and consequently their treatment has more potential side effects [93].

It may be argued that whilst tumours like MB occur at significantly different rates in the adult and paediatric populations, they exhibit the same morphology and hence the distinction between adult and paediatric research is not important. However, findings from translattional oncological research have shown that tumour types with the same morphological characteristics can often have a diverse set of genetic profiles [113]. For the particular example of paediatric MB, it has been recently shown that they are genetically distinct from their adult forms [113]. Consequently, response to anticancer therapy is likely to be different between adults and children. In terms of relevance to this thesis, such variations that occur on a genetic level are likely to translate to variations in textural characteristics; hence, paediatric brain cancer is considered as a separate problem throughout this thesis.

2.4 Summary

This chapter outlined the basic principles and concepts of MR imaging used throughout this thesis. Whilst the technical work presented in this thesis involves the analysis of already acquired data in image space, understanding the physical processes underlying MRI is important. It is particularly important to understand how the choice of certain parameters, such as echo time, affects the contrast characteristics of the imaging data that will be used in our analysis.

Medulloblastoma, pilocytic astrocytoma and ependymoma tumours were then introduced later in the chapter. The focus is on these three types, as they are the most commonly occurring brain tumours affecting childhood. Since the technical aspects of this thesis are based on the analysis of conventional MR data of patients diagnosed with these three tumour types, understanding their MR imaging appearance is of particular importance. Hence, an overview of the manifestation of these tumours on T1 and T2-weighted images was presented.

Chapter 3

Background on Machine Learning

3.1 Introduction

Machine learning is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data [109]. It is an on-going area of development in computer science and blends with parallel developments in statistics, in particular, statistical learning. Accoring to Abu-Mostafa et al [51], the essence of a machine learning problem can be captured in three important components:

- Presence of a pattern.
- Lack of a mathematical means of modelling such pattern.
- Availability of data.

The problem of computer-based classification of brain tumours, using MRI TA, satisfies these three conditions and therefore represents a classical machine learning scenario. In this regard, the primary aim of this chapter is to provide an overview of popular machine learning tools and techniques, particularly those that were needed to achieve the contributions of this thesis.

The chapter starts by providing an overview of different types of learning paradigms. The chapter then provides a discussion on dimensionality reduction, classification algorithms and model evaluation techniques. Since the technical aspects of this thesis are carried out within a *supervised classification* framework, the focus of this chapter is on this type of learning problems.

3.2 The Learning Problem

3.2.1 Aim of a Learning Task

In mathematical terms, the aim of a machine learning task is to use available data to estimate some target function:

$$Y = f(X) + \epsilon \tag{3.1}$$

where:

- X is a member of the feature variables set $X_1, .., X_p$
- Y is an output variable
- ϵ is the random error term

A data sample can be described by the sample pair (X, Y). A *feature* can be defined as *"an individual measurable property of a phenomenon being observed"* [54]. Within the context of this thesis, an example feature would be a textural property that can be measured from an MR image of a tumour, such as histogram's mean grey-level value. An example output variable would be the true diagnosis to which a tumour belongs, for example: medulloblastoma. A tumour sample can therefore be represented by a set of textural features and true diagnosis class.

3.2.2 Types of Learning

Most learning problems fall into one of two categories: *supervised* or *unsupervised*. In supervised learning, for each observation of the feature value(s) x_i, where *i=1,....,n*, there is an associated output response measurement y_i. We wish to fit a model that relates the response to the features, with the aim of predicting the

response in the future, or to better comprehend the relationship between the response and the features [51], [52]. Many classical learning methods such as logistic regression and support vector machines operate in a supervised fashion.

Unsupervised learning is the more challenging situation where every observation i has a vector of feature values x_i but no associated response label y_i [51], [52]. One technique that can be used in this setting is clustering, where the goal is to establish, on the basis of available features, whether the observation falls into one of relatively distinct groups. Unsupervised learning is, however, beyond the scope of this thesis, and the rest of this chapter is focused on supervised settings.

A machine learning task could either be a *classification* or a *regression* learning problem. In order to understand the difference between classification and regression learning problems, it is important to appreciate *quantitative* and *qualitative* variables. Quantitative variables assume numerical values, for example, a person's age, height or income. Within the context of this thesis, quantitative variables are textural parameters extractable from MR images of brain tumours e.g, histogram statistics and grey-level co-occurrence matrix features, which will be explained in detail in the next chapter. In contrast, qualitative variables assume values that fall in one of K different classes, such as a person's gender. Tumour diagnosis (e.g, medulloblastoma, pilocytic astrocytoma, ependymoma) is a qualitative variable. The machine learning community tends to refer to learning problems with a qualitative response as *classification* problems, while those involving a quantitative response as *regression* problems. Whether the response is quantitative or qualitative forms the basis of selecting learning methods, however, whether the features used are qualitative or quantitative is a less important question.

In this regard, the problem of computational classification of brain tumours from MRI textural features can be described as a *supervised classification problem.*

The rest of this chapter discusses some of the most important techniques and concepts that arise in carrying out a supervised classification task.

3.2.3 The Supervised Classification Framework

There are four main steps that comprise a supervised classification experiment:

- First, **feature extraction** is carried out on a set of measured data and builds derived values that aim to capture representative patterns of the data. Such features should ideally be informative and non redundant. In this thesis, features are extracted from MR imaging data using texture analysis (TA) techniques. Such techniques are explained in Chapter 4.

- Secondly, **feature selection and dimensionality reduction** can be carried out with the aim of identifying and removing as many irrelevant and redundant features as possible from the dataset. An irrelevant feature does not affect the target concept in any way, and a redundant feature does not add anything new to the target concept [58]. Irrelevant and redundant features are problematic because they may confuse the learning algorithm, by helping to obscure the distributions of the subset that holds influential features [58],[59].

- A learning algorithm is then **trained** using labelled data in order to build a model that aims to capture patterns, which can accurately classify future unseen data points.

- Finally, the model needs to be **validated and evaluated** using appropriate evaluation metrics and techniques to ensure efficacy and robustness.

The rest of this chapter provides a discussion on feature selection, classification and validation techniques that are used in the technical aspects of this thesis.

3.3 Feature Selection and Dimensionality Reduction

Feature selection is a data pre-processing stage generally carried out prior to the application of learning algorithms, with the aim of identifying and removing as many irrelevant and redundant features as possible from the dataset [58], [59]. In addition to this, another motivation for carrying out feature selection tends to be interoperability. Within the context of this thesis, looking for reasons why particular features train certain algorithm well, and analysing their biological meaning, would be a rather challenging task to be performed on the entire feature set. Reducing the required number of features to a manageable number would enable better analysis and understanding of the model's behaviour.

Taking the example of features extractable via texture analysis, which will be thoroughly discussed in Chapter 4, the use of all possible combinations from all techniques, modalities and datasets would give a very large number of features: over 560 for 3D features! To address this, the technical aspects of this thesis make use of a number of feature selection techniques, which are explained below.

One method that can be used for feature selection is *ReliefF* [56]. The idea behind this technique is to estimate the effectiveness of a feature based on how well its value compares to its neighbours. This is done by searching for an instance's nearest neighbours and finding an instance from the same class (nearest hit) and another one from a different class (nearest miss). The algorithm then uses a weighting approach to estimate the quality for each feature. Good features are assumed to have the same value for instances from the same class and should differentiate between instances that belong to different classes.

Another technique that can be used for feature selection is *entropy minimum*

descriptive length (MDL) [57]. Entropy-MDL is conventionally a feature discretisation technique that works by finding a splitting value (cut-off point) which yields the best gain in entropy, allowing continuous feature values to be discretised. This is repeated recursively with a stopping criteria that is based on the minimal description length (MDL) principle [1] [103],[104]. The basic idea behind MDL principle is to equate data compression with finding regularities in the data. The use of entropy MDL technique as a feature selection method is based on the assumption that since a feature's entropy can be used as a measure of its discriminative power, those features that were rejected by the algorithm can be assumed to be redundant. A feature would not be discretised if no appropriate cut-off points are found [58].

Feature selection falls under a broader category referred to as dimensionality reduction. While feature selection methods work by deciding on a limited number of features to be included in the final sub-set, other dimensionality reduction techniques work by mathematically transforming the data to a new space of fewer dimensions. A popular example is *principal component analysis* (PCA).

PCA aims to reduce the dimensions of a $n \times p$ data matrix X by linearly transforming the data to identify dimensions of maximum variation within the matrix. The data is transformed into a space spanned by a set of orthogonal vectors called the *principal components* (PCs), which are aligned along the axes of maximum variation [86],[105]. The first PC is the dimension with maximum variation and each further PC corresponds to less variation than the previous. In classification settings, a common assumption is that the dimensions of maximum variation, the PCs, are also the dimensions which are most important for classification [86]. If

[1] The minimum description length (MDL) principle is a formalisation of Occam's razor, in which the best hypothesis for a given set of data is the one that leads to the best compression of the data.

this assumption is true, discarding the PCs that correspond to the smallest variance should not cause any degradation in classification performance. This is the basis of using PCA for dimensionality reduction.

3.4 Classification Methods

In this section, we discuss common classification techniques that are used in the technical aspects of this thesis.

3.4.1 The Bayes Classifier

One method that can be trained very efficiently in a supervised learning setting is the Bayes classifier. This classifier works by simply assigning each observation to the most likely class j, given a particular feature x_0, as per its conditional probability [52], [99]. In this context, conditional probability is the probability that Y=j, given x_0 , as per :

$$Pr(Y = j|X = x_0) \tag{3.2}$$

The Bayes classifier will always choose the class for which the conditional probability is largest, hence, the error rate at X = x_0 will be:

$$1 - \max_j Pr(Y = j|X = x_0) \tag{3.3}$$

This is called the *Bayes error rate*, and is probability that the classifier incorrectly classifies an instance.

The above description assumes the need to predict an outcome given only one piece of evidence (i.e. one feature). In practice, classification is done on the basis

of values of multiple features. One approach this could be tackled is by uncoupling the available features. In other words, all features are assumed to be conditionally independent given the class label. Even though this is usually false, the resulting model is easy to fit and works surprisingly well [99]. This approach is referred to as the *naive Bayes* classifier.

Within the context of this thesis, the probabilistic nature of Bayesian classifiers is particularly advantageous for implementation in computer-aided decision support settings. Using probability distributions, it is straightforward to charatcerise confidence in classifier predictions, allowing predictions with low confidence to be rejected and passed on to human experts.

3.4.2 k-Nearest Neighbours Classifier

The k-Nearest Neighbours (kNN) classifier is intuitively very attractive: assign to the data point that is to be classified the class label of the nearest k known data points, where *nearest* is some distance metric such as Euclidean distance [100].

Given

- a positive integer k, and
- a test observation x_0.

kNN first identifies the k points in the training data that are closest to x_0, represented by N. kNN then estimates the conditional probability for class j, as the fraction of points in N, whose response values equal j [52]:

$$Pr(Y = j|X = x_0) = 1/k \sum_{i \epsilon N} I(y_i = j) \tag{3.4}$$

Where I is an indicator variable that assumes the value of 1 if $y_i = j$ is true, and zero if it is false, for the ith data point.

Finally, kNN applies Bayes rule and classifies the data point to the class with the largest probability.

It is important to note that the choice of k can have serious effects on the classification performance. With very small k values, such as 1, the classifier's decision boundary is overly flexible, yielding a classifier that has low *bias* but very high *variance* [52]. As k grows, flexibility decreases and the decision boundary becomes so close to linear. This yields a low-variance but high-bias classifier [52] In machine learning, variance gives an indication to the amount by which $\widehat{f}$ would change if it was estimated using a different data set. Ideally, the estimate for f should not vary too much between training sets. However, a high variance method means that a small change in the training data can result in large changes in $\widehat{f}$. On the other hand, bias refers to the error that is introduced by approximating a complicated problem by a much simple model [52], [108].

Contrary to other learning methods such as SVM, which will be introduced later in this section, kNN uses all available training data for classifier construction once k is defined; there is no summarisation or discard of data points [84]. This can be an advantage, as less frequently occurring classes with unusual characteristics are not ignored. On the other hand, this can be disadvantageous if the training data is noisy or falsely labelled. Hence, it might be expected that a kNN classifier can perform well for relatively rare brain tumour types, such as ependymoma.

3.4.3 Classification Tree

Classification tree algorithms, in their simplest form, are hierarchal If-Else statements that can be applied to predict a result based on available data. When building learning models, classification trees are a good choice when the goal is to generate classification rules that can be easily comprehended [52]. Within the

context of this thesis, the use of a tree-based classifier to diagnose tumour classes from MR images may be appealing to clinicians as it would be a straightforward approach for translating unclear textural patterns to logical structures. Trees can be built through a process known as *binary recursive partitioning*. This is an iterative task that requires categorising the data into partitions, and then splitting the partitions further on each of the branches.

In broad terms, the process initiates at a single node, followed by looking for the binary distinction which gives us the most information about the classes, as measured by *information gain* [101]. The same is done on each of the resulting nodes, in a recursive manner, until a pre-defined stopping criterion is reached [101]. This is detailed in Algorithm 1. To calculate information gain, *entropy* - a measure of uncertainty of a random variable - is calculated first [107]:

$$H(X) = -\sum_{i} Pr(x_i) \log Pr(x_i) \tag{3.5}$$

and the entropy of X after observing the values of another variable Y is defined as:

$$H(X|Y) = -\sum_{i} Pr(x_i) \sum_{i} Pr(x_i|y_i) \log Pr(x_i|y_i) \tag{3.6}$$

where:

- $Pr(x_i)$ is the prior probabilities for all values of X
- $Pr(x_i|y_i)$ is the posterior probabilities of X given the values of Y.

Information gain can finally be calculated as follows:

$$InformationGain(X|Y) = H(X) - H(X|Y) \tag{3.7}$$

Algorithm 1 An algorithm describing the recursive process of building a decision tree [101].

```
T ← EmptyTree
if all instances in D have the same class c then
    label(T) ← c return T
else
    if features = 0, or no feature has positive information gain then
        label(T) ← MostCommonClassInD return T
    end if
    A ← FeatureWithHighestInformationGain
    label(T) ← A
    for each value a of A do
        D_a ← InstancesInDwithA = a
        if D_a = 0 then
            T_a ← EmptyTree
            label(T_a) ← MostCommonClassInD
        elseAddBranchFromTtoT_aLabelled(a)
        end if
    end for
end if
```

3.4.4 Logistic Regression

Logistic regression[2] models the probability of outcome Y given a certain feature X, $Pr(Y = 1|X)$. Then, for any given value of X, one might predict whether a given class label Y is positive or negative (assuming a binary classification problem). For example, for any data point where Pr(X) $\geq$ 0.5, Y gets classified as positive. This probability can be estimated using the *logistic function*:

$$Pr(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \tag{3.8}$$

[2]This is called logistic regression due to its similarity to linear regression; although it is a form of classification, not regression.

where β_0 and β_1 are the model parameters.

By manipulating Eq 3.8, we find that:

$$Pr(X)/(1 - Pr(X)) = \exp(\beta_0 + \beta_1 X) \tag{3.9}$$

By taking the logarithm of both sides of Eq 3.9, we arrive at

$$\log(Pr(X)/(1 - Pr(X))) = \beta_0 + \beta_1 X \tag{3.10}$$

The left-hand side of this relationship is referred to as the *log-odds*, or *logit*. One can estimate β_0 and β_1 using a technique called *maximum likelihood* [52]. The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we try to find estimates for β_0 and β_1 such that the predicted probability Pr(X) yields a number close to one, for all data points that fall into the positive class, and a number close to zero for all data points that fall into the negative class. This intuition can be mathematically formalised using the maximum likelihood function l, as shown in Eq 3.14:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} Pr(x_i) \prod_{i':y'_i=1} 1 - Pr(x'_i) \tag{3.11}$$

3.4.5 Artificial Neural Network

Artificial Neural Network's (ANN) name has its origins in attempts to find mathematical representations of information processing in biological systems [54]. In biology, a neural network consists of a large number of nerve cells, *neurons*, which are basic signalling units with several inputs but one output. Depending on the neuron's location, its inputs and outputs may be from/to sensory organs/motor nerves, or other neurones within the network. In broad terms, each input to the neuron connects through a synapse, which controls the gain of the signal from each source.

When designing an ANN, the basic unit that mimics the behaviour of a neuron is a *perceptron*. The perceptron works by taking several inputs with their associated weights, and depending on whether the combined input weight exceeds a pre-defined threshold, a certain output signal will be activated (Fig 3.1 (a)). The *perceptron* can be mathematically described as shown in Equation 3.15:

$$y = \phi\{\sum_{i=1}^{n} w_i x_i + b\} \tag{3.12}$$

Where:

- y is the output signal,
- ϕ is the activation function, which translates the input signals to output signals. Common types include unit step, sigmoid and Gaussian [102].
- n is the number of connections to the perceptron,
- x_i is the value of the ith connection,
- b represents the threshold.

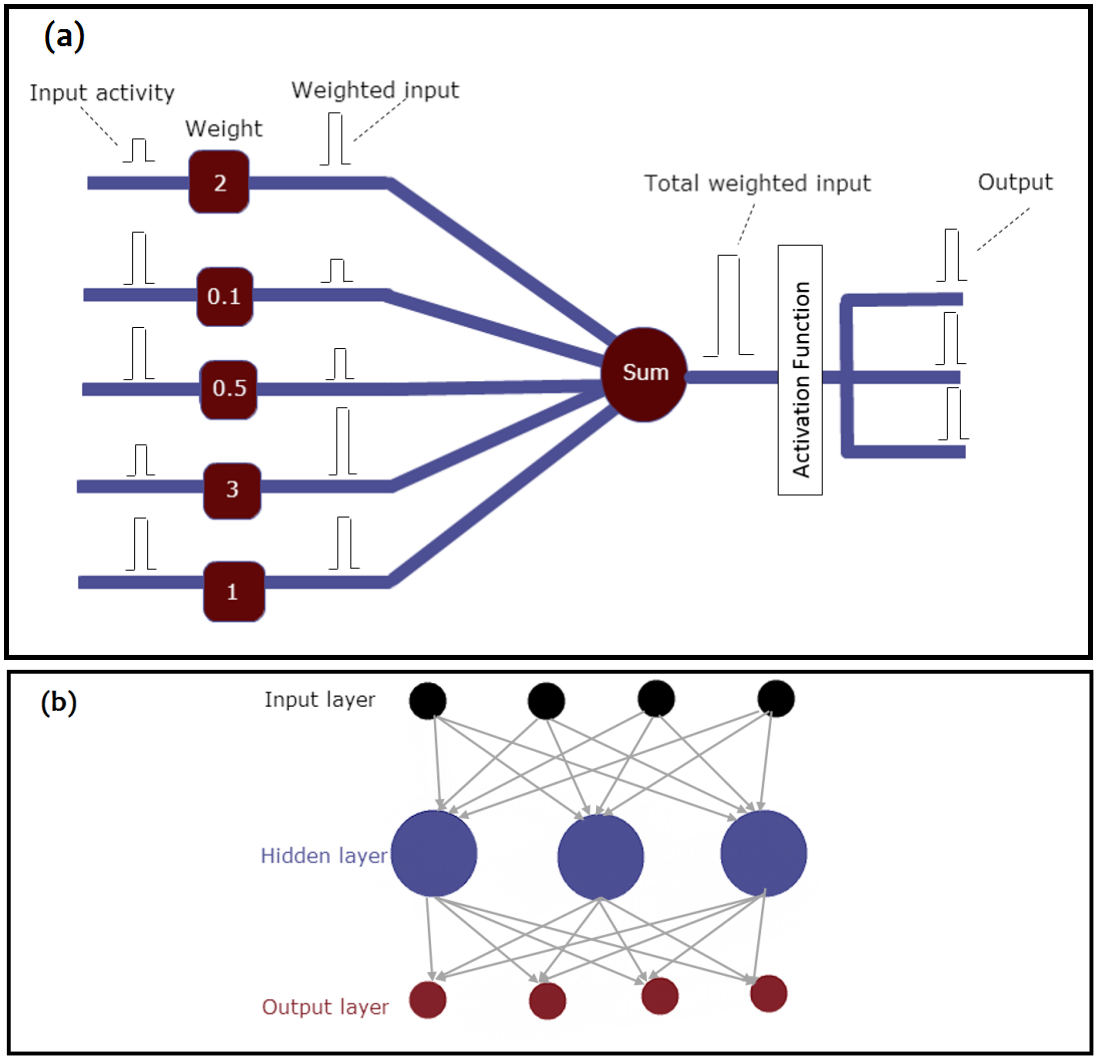


Figure 3.1: A figure showing (a)the idealisation of a perceptron. Each activity is multiplied by a weight and the weighted inputs are then added. The output activity is computed using an activation function (b) an ANN consisting of three layers that are fully connected.

Whilst the design of a single perceptron is simple, its strength can be shown when several perceptrons are combined to work together to form an ANN. A single perceptron is only capable of describing linear separations between data classes, whereas an ANN can describe non-linear regions.

In an ANN, perceptrons are organised in layers, where each layer takes input from the previous, applies weights and then signals to the next layer if appropriate. The use of hidden layers within an ANN alleviates the limitations of using an individual perceptron for learning [86]. To train an ANN to perform some task, the weights of each node must be adjusted in a way that the error between the desired output and the actual output is reduced. This process requires the ANN to compute the error derivative of the weights [87]. In other words, it must calculate how the error changes as each weight is increased or decreased slightly. There are several ways of doing this, most of which involve initialising the weight and feeding the network with an example. The error made by the network at the output nodes is calculated, and fed backwards through a process called *back-propagation*. By repeating this process, the network can update the weights and learn to accurately distinguish between different classes. An example ANN is shown in Figure 3.1 (b).

3.4.6 Support Vector Machines

Support vector machine (SVM) is a classification method that has grown in popularity within the machine learning community since the 1990s. In order to appreciate the mechanisms underlying SVMs, a number of concepts, such as *hyperplanes*, *maximal margin classifiers* and *support vector classifiers* are introduced below.

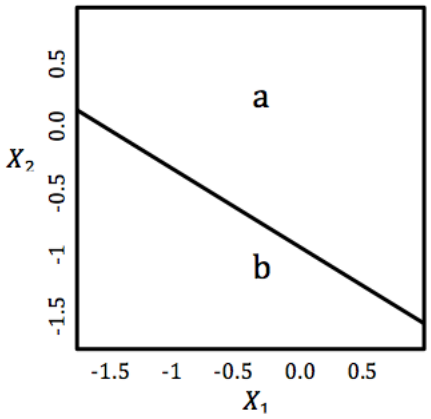


Figure 3.2: A graph showing a one-dimensional hyperplane $1 + 2X_1 + 3X_2 = 0$. Region a is the set of points for which $1 + 2X_1 + 3X_2 > 0$, whereas region b is the set of points for which $1 + 2X_1 + 3X_2 < 0$.

Hyperplanes: In a p-dimensional space, a hyperplane is a flat subspace of dimension p-1. For instance, in two dimensions, a hyperplane is a flat one-dimensional subspace, i.e. a line, as shown in Fig 3.2. A hyperplane is mathematically defined using the simple Equation:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0 \tag{3.13}$$

For parameters $\beta_0, \beta_1, ..., \beta_p$

Consider Figure 3.3 , which shows a number of data points that fall into one of two possible classes. Three separating hyperplanes, out of many possible, are shown as straight black lines in the figure. The situation depicted raises an interesting question: *given a number of possible separating hyperplanes that can per-*

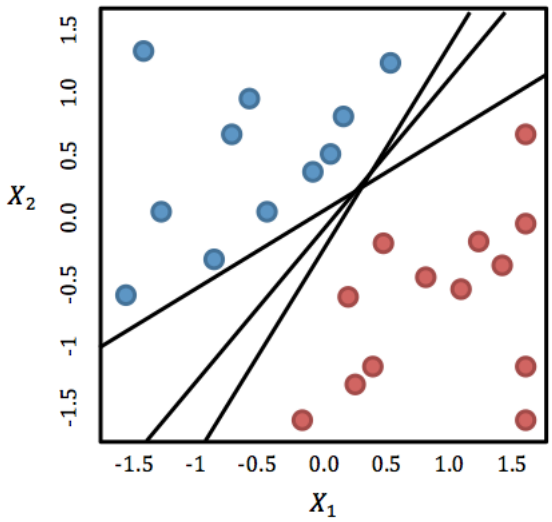


Figure 3.3: A graph showing a number of data points that fall into one of two possible classes, illustrated here in blue and red. Thee separating hyperplanes, out of many possible, are shown as straight black lines.

fectly separate our data points that belong to different classes, is there a reasonable way to decide on only one hyperplane to use?

Maximal Margin Classifier: An intuitive choice is the *maximal margin hyperplane*, which is the separating hyperplane that is furthest away from the training data points. This can be determined by calculating the perpendicular distance from each training point to a given hyperplane; the smallest of which is the minimal distance from the data points to the hyperplane, and is referred to as the *margin*. This is the basis of the maximal margin classifier: a technique that simply classifies a data point based on which side of the maximal margin hyperplane it lies. By inspecting Fig 3.4, one can see that two data points are equidistant from the maximal margin hyperplane and lie along the dotted lines that indicate the margins width. Such data points are referred to as support vectors, as they are vectors that support the maximal margin hyperplane, meaning that a slight shift in their position would cause a shift in the hyperplane.

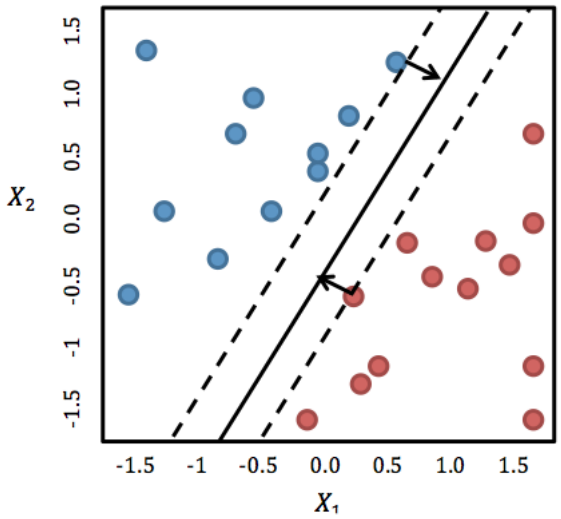


Figure 3.4: A plot illustrating a maximal margin hyperplane separating data points that belong to two classes. The two points that lie on the dashed lines are the support vectors. The arrows indicate the distance from the support vectors to the *hyperplanes* margins.

Support Vector Classifier: Whilst the maximal margin classifier is a straightforward way of classifying data when a separating hyperplane is present, in many cases, it is not possible to construct such hyperplane. This motivates the extension to a concept called *soft margin*, which almost separates the classes by misclassifying a few training observations with the aim of performing better when grouping the remaining data points (figure 3.5). The soft margin forms the basis of a *support vector classifier*: a technique that works by allowing some data points to be on the incorrect side of the margin, instead of seeking the largest possible margin that separates all available training points. Not only can a data point be on the incorrect side of the margin, but also on the incorrect side of the hyperplane.

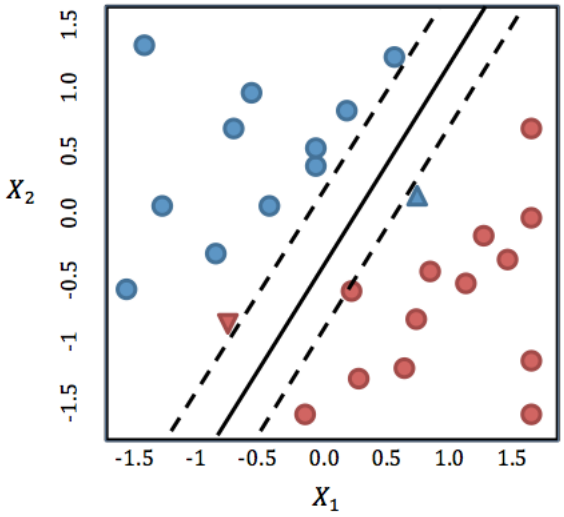


Figure 3.5: A plot illustrating the construction of a soft margin that allows two data points to be on the incorrect side of the hyperplane and margin, but performs well when separating the rest of the data points. The two points on the incorrect side of the hyperplane are plotted as triangles.

In mathematical terms, the support vector classifier is the solution to the optimisation problem:

Maximise M

Subject to $\sum_{j=1}^{p} \beta_j^2 = 1$

$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}) >= M(1 - \epsilon_i)$

$\epsilon_i >= 0, \sum_{i=1}^{n} \epsilon_i <= C,$

where:

- M is the width of the margin.
- C, the cost coefficient, is a non-negative tuning parameter that bounds the sum of the $\epsilon_i's$ and hence determines the severity of margin and hyperplane violations.
- ϵ_1,..,ϵ_n are called *slack variables*.
- x_1,..,x_n are training observations.

- y_1,..,y_n are the associated class labels observations, which are of a binary outcome.
- p is the number of dimensions in the space.

Slack variables allow data points to be on the wrong side of the margin or hyperplane. The slack variable ϵ_i indicates where the ith data point is located, with respect to the hyperplane and the margin. If ϵ_n=0 , then the data point is on the correct side of the margin, whereas if $\epsilon_n > 0$, then it is on the wrong side of the margin (that is, the data point has violated the margin). If $\epsilon_n > 1$, then it is on the wrong side of the hyperplane.

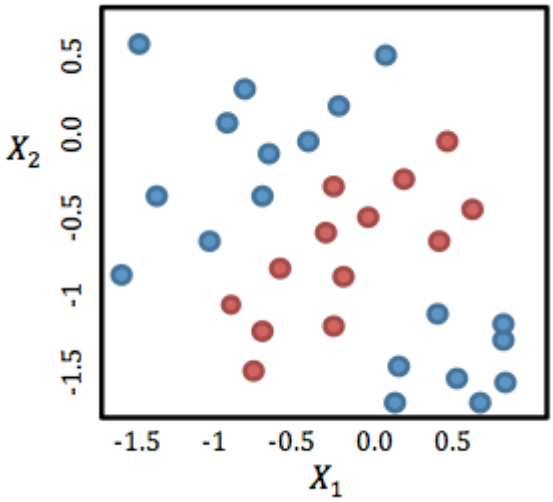


Figure 3.6: A plot illustrating two classes of data that are not linearly-separable.

Support Vector Machines: Thus far, the use of hyperplanes has been discussed within the specific context of classes that are linearly separable. However, datasets used in practice often include non-linearly separable observations (example Figure 3.6). This is certainly the case for textural features extracted from the clinical MR datasets used in the technical aspects of this thesis [37]. SVM addresses this problem by extending the aforementioned support vector classifier to manipulate

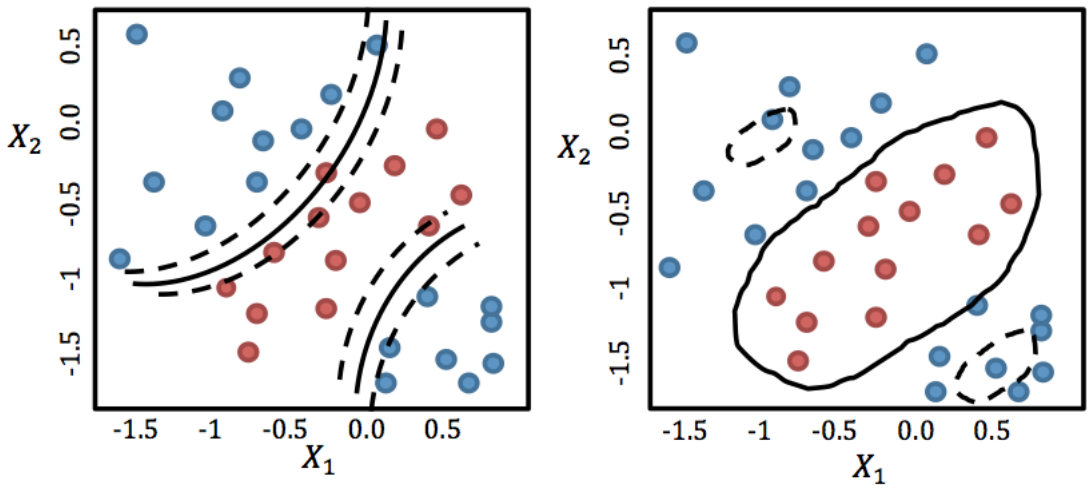


Figure 3.7: A plot illustrating thow polynomial (left) and radial (right) kernels perform on data.

the feature space in a certain way, using *kernels*, as detailed below.

A kernel is a function K that quantifies the similarity between two data points, x_i and x_i', in a p-dimensional space. For instance, we could take:

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{3.14}$$

which characterises the classic support vector classifier. Eq 3.14 is a *linear kernel* due to the linear nature of the support vector classifier. This kernel essentially uses Pearson correlation to quantify the similarity between a pair of data points. This could be extended as:

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d \tag{3.15}$$

which yields a *polynomial kernel* of degree d. The use of a kernel with a degree

higher than 1 essentially leads to fitting a support vector classifier in a higher-dimensional space of degree d, rather than in the original feature space. This yields a more flexible decision boundary when fitting the data points. When a support vector classifier is combined with non-linear kernels such as Eq 3.15, the resulting classifier is an SVM. Another popular kernel is the *radial* kernel, which is characterised as:

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \tag{3.16}$$

where γ is a positive constant. Figure 3.7 shows SVMs with a polynomial kernel (left) and a radial kernel (right) applied to non-linear features and resulting with highly flexible decision boundaries.

SVMs with Multiple Classes: Thus far, the discussion of SVM assumed binary-class situations. Given that this thesis concerns the classification of more than two tumour types, extending SVM to work under such settings is important. However, the concept of hyperplanes does not naturally lend itself to multi-class problems. The two most popular ways to address this are *one-versus-all* and *one-versus-one* classification schemes. Given the problem of classifying MB, PA and EP, one-versus-all assumes that the class of interest (say, MB) is positive while the rest of the classes collectively represent a negative class. This is then repeated for the rest of the classes (PA versus non-PA; EP versus non-EP). Under the one-versus-one scheme, a series of SVMs is constructed, each comparing a pair of original classes (e.g. MB versus PA, MB versus EP, PA versus EP). We then separately classify each test observation using each classifier. The final classification is performed by assigning the test observation to the class to which it was most frequently assigned in these pairwise classifications.

3.5 Model Validation and Evaluation

3.5.1 Measures of Classification Performance

Suppose that we seek to estimate f on the basis of data points that can be used to train, or teach, a learning method. The data points can be represented as (x_1, y_1),.., (x_n, y_n), where y_1, ..., y_n are qualitative class labels. The most straightforward metric for assessing the performance of our estimate is the *training error rate*, which is proportion of mistakes that are made if we apply our estimate f on the same data points used to train the classifier. This can be mathematically described as follows:

$$1/n \sum_{i=1}^{n} I(y \neq \widehat{y}) \tag{3.17}$$

where:

- $\widehat{y}$ is the predicted class label for a particular data point using $\widehat{f}$.
- $I(y \neq \widehat{y})$ is an indicator variable that assumes the value of 1, if $y \neq \widehat{y}$ is true, and zero, if it is false for the ith data point [52].

We are, however, not interested in how the classifier performs on seen data, but rather in the error rate that results from testing our classifier on unseen data points that were not used in the training. A basic rule in classification experiments is that class predictions are not made for data samples that are used for training, because that would lead to an over-optimistic bias of classification performance [100]. The test error is therefore calculated as follows:

$$Ave(y \neq \widehat{y}) \tag{3.18}$$

Where $\widehat{y}$ is the predicted class label that results from applying the classifier to an unseen data point. Instead of computing the error rate, an alternative is to calculate the *classification accuracy*, which is the fraction of *correct*, rather than incorrect, classifications. In the technical parts of this thesis, the classification accuracy metric is extensively used, alongside other measures of classification performance, to assess our learning models. However, accuracy is not always adequate when evaluating classifiers, so other measures that are commonly used in practice are introduced next.

		True Class: **Positive**	True Class: **Negative**
Hypothesised Class	**Positive**	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

Figure 3.8: A diagram showing an example confusion matrix.

Assuming a binary classification problem, and given a classifier and a data point, the point can either be *positive* or *negative*. That is, if we are predicting the presence of a particular type of cancer (say, Medulloblastoma), positive would indicate the presence of Medulloblastoma, while negative would indicate otherwise. Hence, there are four possible outcomes upon classification. If the data point is positive and it is classified as positive, it is counted as a true positive (TP); if it is classified as negative, it is counted as a false negative (FN). If the data point is negative and it is classified as positive, it is counted as a false positive (FP); if it is classified as negative, it is counted as a true negative (TN). A two-by-two confusion matrix can be constructed (Figure 3.8) to summarise these outcomes; this matrix forms the basis of popular classifier evaluation measures [55].

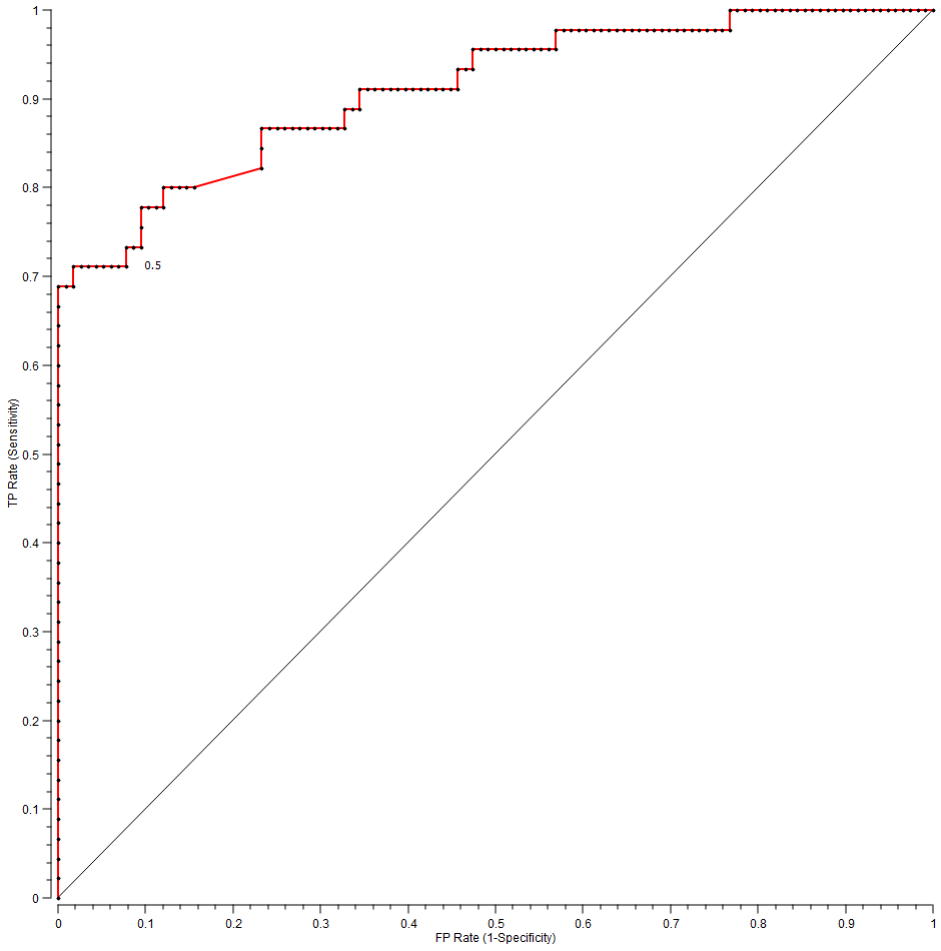


Figure 3.9: A diagram showing an example receiver operator characteristics (ROC) curve.

One measure that can be calculated from a confusion matrix is *sensitivity*. This is the proportion of actual positives that are correctly identified as such. In other words, it is the ratio between true positives and the sum of true positives and false negatives. Another statistical measure that can be calculated is *specificity*, which is the proportion of negatives that are correctly identified, as such. Specificity is the ratio between true negatives and the sum of false positives and true negatives [55].

One way of visually depicting a classifier's performance is the *receiver operator characteristics (ROC) curve*. To plot an ROC curve, one must first calculate the true positive rate (ratio between true positives and all positives) and false positive rate (ratio between false positives and all negatives) for different discrimination thresholds used by the classifier. An example ROC curve is illustrated in Fig 3.9. The lower left point (0,0) represents the strategy of never issuing a positive classification, whereas the top right point (1,1) represents the opposite strategy of unconditionally issuing positive classifications, in terms of true positive and false positive rates. A perfect classification would be identified in point (0,1). The diagonal line $y=x$ represents the strategy of randomly guessing a class. Generally speaking, one point is considered better than another in ROC space if it is to the northwest of the first (higher TP rate, lower FP rate, or both) [55].

Although the ROC curve is a visual depiction of classification performance, it is possible to reduce ROC performance to a single scalar value summarising expected performance, namely *area under the ROC curve* (AUC). AUC is a portion of the area in unit square, and will therefore have a value between 0 and 1. Note that since random guessing (line y=x) produces an area of 0.5, a successful classifier must have an AUC value greater than 0.5.

3.5.2 Model Validation Schemes

Due to limited cohort sizes, having an independent testing set is not always possible, which motivates the need to use other means for measuring a classifier's generalisation performance. This is particularly relevant to this thesis due to the limited cohort sizes, typical in paediatric neuro-oncology.

k-fold cross-validation (CV) provides an easy way evaluate a classifier's performance. Here, the dataset is partitioned into k folds, and each cross-validation loop involves the use of k-1 folds for training and the remaining data for testing. This process is repeated k times, which ensures that each of the subsets is used once for testing, and the results are averaged over the folds [100],[106].

A variant of k-fold CV is *stratified CV*. This approach makes sure that folds are selected so that the mean response value is approximately equal in all of them. This ensures that each fold contains roughly the same proportions of the available class labels, and is known to produce results with a lower variance than the conventional approach. Another variant of k-fold cross-validation, and perhaps the most popular one, is the *leave-one-out cross-validation* (LOOCV) scheme. Here, k is equal to the number of observations, and the test fold has only one element [100],[106].

Alternatively, one could use the *repeated random sampling* approach. This requires random partitioning of the dataset into a training and a testing set of fixed sizes (say 75% for training and the remaining for testing). The partitioning is then repeated and the results are averaged over all repetitions [106].

3.5.3 The Problem of Over-fitting

When training a learning algorithm in a supervised classification experiment, the aim is to use the available features to build a classification model that can gener-

alise well with the data, i.e. apply to the larger population from which the training sample was drawn [110]. The over-fitting problem arises when the model incorporates error or noise from the training sample and consequently does not generalise well to the overall population [110]. In other words, the chosen model may have a large out-of-sample error, despite demonstrating a low in-sample error [51].

Over-fitting is likely to occur when the number of features in the model is larger relative to the size of the training set [110]. The rarity of paediatric brain tumours and the abundance of available textural features that can be extracted from MR images means that over-fitting will be a potential problem in the experimental parts of this thesis. This can, however, be mitigated through the use of dimensionality reduction techniques to reduce the feature to sample size ratio. Additionally, the use of cross-validation can address the lack of sufficient training sets.

3.5.4 The Problem of Class-Imbalance

Another potential problem that is likely to arise in the technical aspects of this work, and is linked to the rarity of paediatric brain tumours, is the issue of class-imbalance. Imbalanced data usually refers to classification problems where available data points that belong to different classes are not equally represented. Given that some tumour types are a lot more frequently occuring than others (e.g. MB versus EP), our cohort is likely going to exhibit this issue. The machine learning community has tackled the class imbalance issue in two ways. One is to assign costs to training examples, and the other is through sampling the data (by over-sampling minorities or under-sampling majorities)[73]. Given the already limited cohort sizes in paediatric oncology, under-sampling majority classes is an unlikely choice for addressing this issue, within the context of this thesis.

3.6 Summary

Since much of this thesis is devoted to the use of textural features within a supervised classification framework, it was necessary to review common machine learning techniques that were applied as part of the work presented. An overview of different types of learning paradigms was presented. This was followed by an explanation of a number of feature selection and dimensionality reduction techniques. A review of common classification algorithms (Naive Bayes, Classification Tree, Logistic Regression, kNN, ANN and SVM), which were used in the technical parts of this thesis, was then given. Finally, the chapter concluded with a summary of techniques and metrics that can be used to assess the performance of learning models. The next chapter follows by introducing TA techniques that will be used to extract features from MR imaging datasets.

Chapter 4

Texture Analysis of MR Images: Theory and State-of-the-Art

4.1 Introduction

This chapter gives a background on texture analysis of MR images, with a particular focus on its application as a tool for brain tumour characterisation. The first section of the chapter introduces theoretical concepts behind texture analysis, with an emphasis on statistical methods that aim to analyse grey-scale images. MRI scans are grey-scale, hence the focus on this category of digital images in this chapter. The material presented in the second section provides an extensive literature review of the current state-of the-art for characterising paediatric and adult brain tumours.

4.2 Background on Texture Analysis

4.2.1 Introduction

Figure 4.1: Different types of image textures that human vision processes on a daily basis. Original image was obtained from [15].

The concept of texture is generally regarded as a broad one, with no standard definition that is universally agreed upon. This is perhaps due to the numerous interpretations in which humans perceive texture [16]. The Oxford dictionary defines texture as: *"the feel, appearance or consistency of a surface or a sub-*

stance" [17]. In this thesis, the main focus is on visually sensed textures. Figure 4.1 shows examples of different types of textures that human vision can encounter on a daily basis. The figure is illustrative of the level of sophistication that the human brain deals with when processing visually sensed textural patterns. The difference in textural properties, for instance, defines the borders between clear sky and clouds; such variations in textural patterns allows us to visually distinguish between different objects.

In the field of medical image analysis, texture is defined as *"the spatial variation of pixel intensities within an image"* [18]. Textural features are mathematically defined parameters computed from pixel distributions and intensities, and can be used to characterise the surface of a given object. This means that textural features can provide a quantitative description of the spatial information contained within an image, and can therefore be useful in a variety of applications [18],[19]. Texture analysis (TA) is the term used for methods that can compute such features [20].

Although research in TA has been of interest since the 1970s, the performance of TA-based techniques was greatly hindered in the past by the difficulty of acquiring high-quality imaging data. It is due to the recent advances in digital imaging that TA has become of great potential across a range of applications [21], such as MRI. Early work on TA focused on ways of quantifying geographical information obtained from aerial photographs and satellite images [22]. With improvements in digital imaging, TA found its way to a variety of applications that ranged from biometric identification systems [23] to content-based image retrieval [24]. MR imaging is potentially one of the most exciting imaging technologies where TA can be applied, because MRI offers good soft tissue contrast and provides powerful experimental control of the spatial resolution and signal intensity changes during the imaging experiment [21]. With regards to the medical MR imaging

literature, TA was successfully used for image characterisation in a wide range of disorders and diseases, including brain cancer [25], breast cancer [26] and multiple sclerosis [27].

According to Castellano et al [28], TA methods can be divided into four main categories: *statistical, structural, model-based and transform.* Statistical methods aim to represent textures using pixel intensities, distributions and relationships. These are particularly common in medical image analysis applications. Structural methods aim to represent textures using well-defined primitive objects e.g. straight lines can be used to represent a square object. Model-based techniques exploit complex mathematical models (e.g. stochastic) to carry out TA, whereas transform methods use techniques such as Fourier Transform or Wavelet Analysis to extract textural features. This thesis is focused on the use of statistical TA methods, the most common of which are detailed below[1].

4.2.2 Common Statistical TA Methods

The techniques discussed below are those used in the technical aspects of this thesis (chapters 5-7).

(a) Histogram Statistics

In grey-scale images, pixel values range between 0 and $2^b - 1$, where b is the disk memory (in bits) occupied by each image pixel. Generally, 8 bits are sufficient, giving grey-level values that range between 0 and 255. However, medical MR images tend to use 12 bits in order to enhance tissue visualisation, giving grey-level values that range between 0 and 4095. Low pixel values are attributed to

[1]Whilst auto-regressive model is extensively used in statistics, particularly time-series analysis, it is considered as a model-based technique in this thesis, as per Castellano et al, for simplicity.

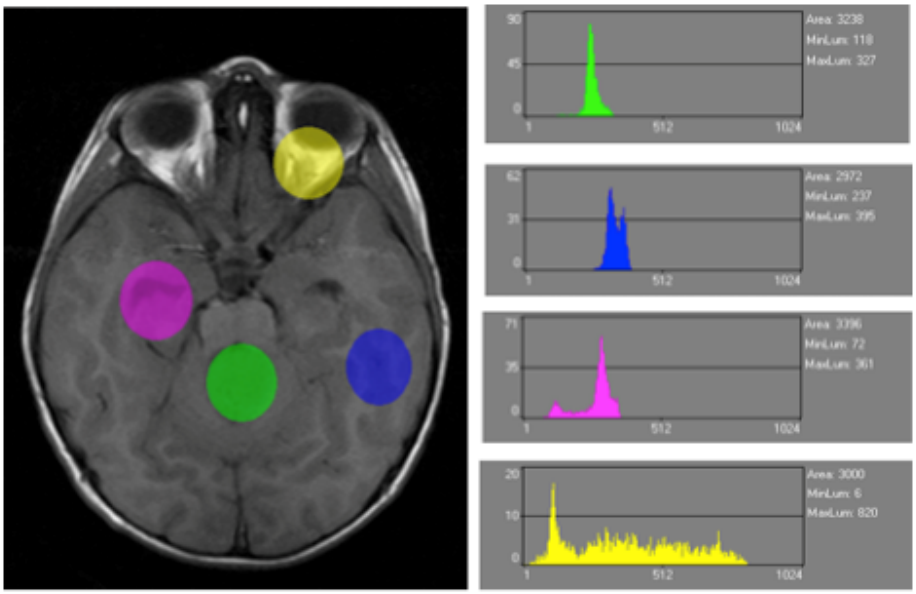


Figure 4.2: Four regions of interest (ROIs) and their corresponding histograms extracted from an axial T1-weighted MR image. The ROI marked in green includes a tumour lesion. Original image was obtained from CCLG database [4].

darker grey-levels and high pixel values are attributed to lighter grey-levels [28].

A grey-scale image histogram is the count of how many pixels in the image possess a certain value. The shape of a histogram provides many clues to the visual characteristics of the corresponding image. For instance, a narrowly distributed histogram indicates that the image is low-contrast and a bimodal histogram suggests that that the image may contain an object with a narrow intensity range against a background of differing intensity [80]. Figure 4.2 shows the histograms of four regions of interest (ROIs) obtained from a T1-weighted MR brain scan.

In order to quantify image properties, a number of features can be extracted from a histogram, namely *mean, variance, skewness and kurtosis*. Equations 4.1 to 4.4 list the features extractable from a grey-level histogram, together with their importance and formulae. Histogram analysis is considered a first-order statistical TA technique, as it does not take into account spatial pixel neighbourhood

relationships or dependencies [80].

Assuming that N_g is the number of distinct grey-levels in an image ($i = 1, 2, ..., N_g$), and that $p(i)$ is the normalised histogram vector (entries are divided by total number of pixels), histogram features can be calculated as follows:

- *Mean*: Measures the average grey-level value of the image

$$\mu = \sum_{i=1}^{N_g} ip(i) \tag{4.1}$$

- *Variance*: Shows how far from the mean the grey-levels are distributed. This gives an idea about how homogeneous the pixel distribution is.

$$\sigma^2 = \sum_{i=1}^{N_g} (i - \mu)^2 p(i) \tag{4.2}$$

- *Skewness*: Measure of the datas lack of symmetry. Data can be described as symmetric if it looks the same to the left and right of the distributions centre point.

$$\mu_3 = \sigma^{-3} \sum_{i=1}^{N_g} (i - \mu)^3 p(i) \tag{4.3}$$

- *Kurtosis*: Measure of whether the data is peaked or flat relative to the normal distribution. Data with high kurtosis tends to have distinct peaks near the mean, decline rapidly and have heavy tails.

$$\mu_4 = \sigma^{-4} \sum_{i=1}^{N_g} (i - \mu)^4 p(i) - 3 \tag{4.4}$$

(b) Absolute Gradient

The gradient of an image is a measure of the local spatial variation of grey-level intensities across the image. For instance, if at a point in an image the grey-level intensity varies rapidly from black to white, the resulting gradient would be high. If the intensity varies from, say, light grey to dark grey, the resulting gradient would be low at that point. Whilst the gradient may be positive or negative, we generally discard the sign, as we are mainly interested in whether the variation is abrupt or smooth. In other words, the main interest is in the gradient's *absolute value* [28]. Figure 4.3 shows a T1-weighted brain MR image and its corresponding gradient image.

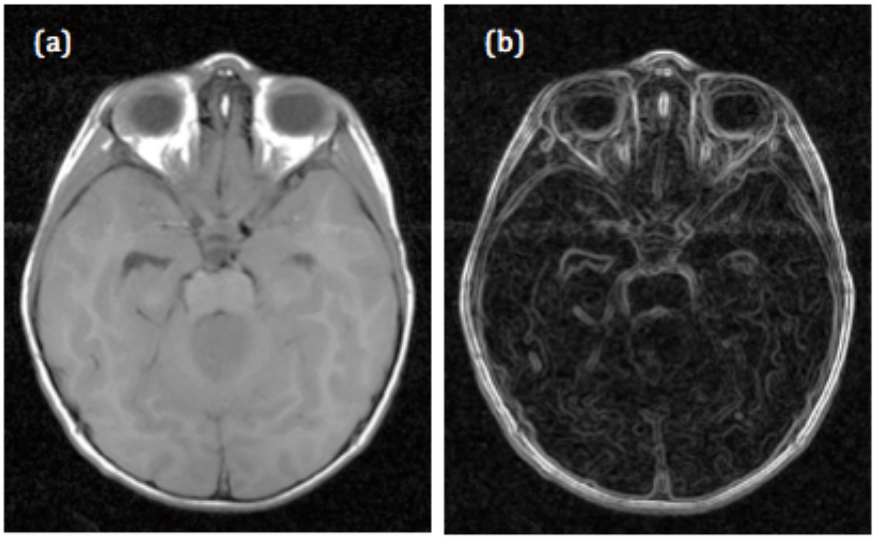


Figure 4.3: (a) An axial T1-weighted MR image and (b) its corresponding gradient image. Original image was obtained from the CCLG database [4].

Figure 4.4: A grey-level image showing a hypothetical pixel neighbourhood.

Assuming the hypothetical neighbourhood for point x(i,j) as per Fig 4.4, the absolute gradient value can be calculated for each pixel as shown in Equation 4.5:

$$AbsGr = \sqrt{(R-H)^2 + (N-L)^2} \tag{4.5}$$

Quantitative features that could be extracted using this technique are mean, variance, skweness and kurtosis, which are estimated from the histogram of the absolute gradient image. Absolute gradient is also considered a first-order statistical technique.

(c) Grey-Level Co-Occurence Matrix

Grey-level co-occurrence matrix (GLCM) is a second-order TA technique that was introduced by Haralick et al [22] and allows for the extraction of statistical information about pixel pairs distribution. In particular, GLCM computes how often a pixel with value i, occurs either horizontally, vertically or diagonally to adjacent pixel with value j.

In order to compute the GLCM of an image, it is necessary to define a distance (d) and a direction (θ) of analysis first (Figure 4.5 (a)). Pixel pairs separated by this distance are then analysed across the specified direction, which is done by counting the number of pixel pairs that assume a certain grey-level sequence. To illustrate how GLCM is computed, assume that we define the direction to be horizontal and the distance to be one pixel. The GLCM element denoted by *p(1,2)* will correspond to the number of pixel pairs that were found in the image which have the values 1 and 2 respectively, and are horizontally separated by one pixel. Figure 4.5(b) shows a hypothetical image and its corresponding GLCM, assuming a horizontal direction of analysis and a one-pixel distance.

Note that it is common to calculate multiple GLCMs for a single image; one for each pair of distances and directions defined. It is usual to use distances that range from 1 to 4 pixels in the horizontal (0^o), vertical (90^o) and two diagonal directions (45^o and 135^o). Several textural features can be derived from the GLCM, most of which aim to quantify the image's *homogeneity (smoothness)* or *heterogeneity (coarseness)* levels. A summary of the features extractable from a GLCM, together with their importance and formulae, is shown in equations 4.6 to 4.14.

Assuming that:

- N_g is the number of distinct grey-levels in the image ($i, j = 1, 2, ..., N_g$), and item $p(i, j)$ is the $(i, j)^{th}$ entry in a normalised GLCM

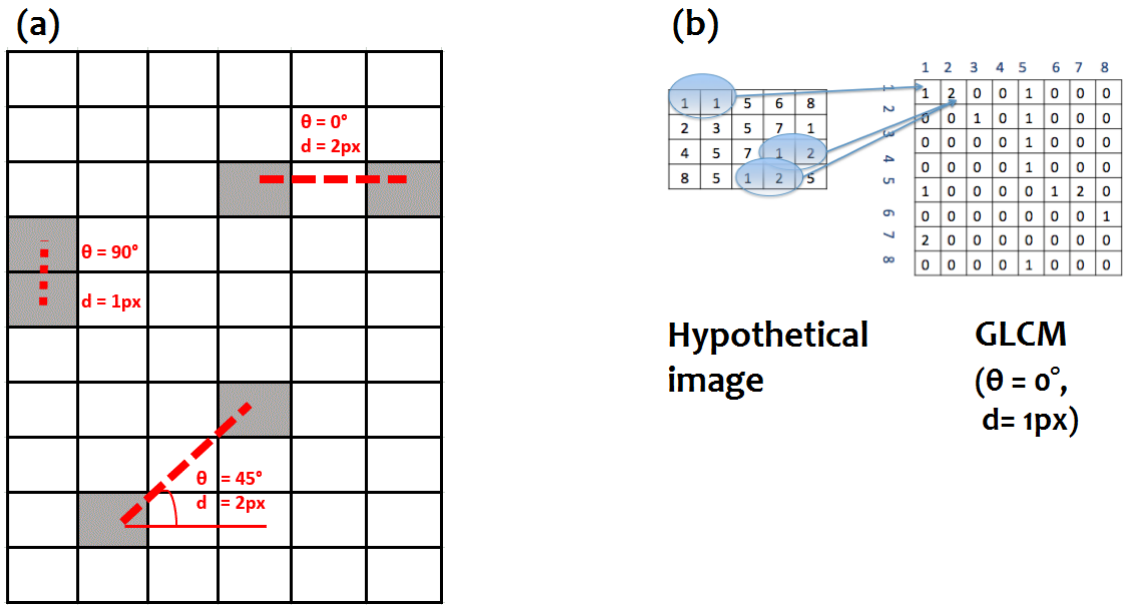


Figure 4.5: (a)Illustration of the pixel relationships considered by the Grey-Level Co-Occurence Matrix (GLCM) technique.(b)A hypothetical image and its corresponding GLCM, assuming a horizontal direction of analysis and a 1-pixel distance

.

- m_x, m_y , σ_x , σ_y are the mean and standard deviation values of rows and column sums of the GLCM respectively, related to the marginal distributions $p_x(i)$ and $p_y(j)$
- $p_x(i) = \sum_{i=1}^{N_{rows}} p(i,j)$
- $p_y(j) = \sum_{j=1}^{N_{column}} p(i,j)$

Then GLCM features can be extracted as follows:

- *Angular Second Moment (ASM)*: Measure of local homogeneity; high ASM values indicate good homogeneity due to the presence of only a few grey-levels, giving a GLCM with only a few but relatively high values of p(i,j).

$$ASM = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i,j)^2 \tag{4.6}$$

- *Contrast (CON)*: Estimates local variation; high CON values indicate low homogeneity.

$$CON = \sum_{n=0}^{N_g-1} n^2 \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i,j)^2 \text{ , where } n = |i-j| \tag{4.7}$$

- *Inverse Different Moment (IDM)*: Additional measure of homogeneity; high IDM indicates a smooth texture.

$$IDM = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i,j)^2 / \{1 + (i-j)^2\} \tag{4.8}$$

- *Entropy (ENT)*: Measure of randomness within the image; high ENT indicates low homogeneity.

$$ENT = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i,j) \log p(i,j) \tag{4.9}$$

- *Correlation (COR)*: Measure of the level of spatial dependencies of grey-levels within the image.

$$COR = \frac{\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (ij) p(i,j) - m_x m_y}{\sigma_x \sigma_y} \tag{4.10}$$

- *Sum of Squares (SSQ)*: SSQ is the variance computed from the GLCM. It is similar to ENT in terms of measuring scatter from the mean.

$$SSQ = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (1-m)^2 p(i,j) \tag{4.11}$$

- *Sum Average (SumAvg)*: SumAvg measures the mean of p_{x+y}

$$SumAvg = \sum_{i=0}^{2N_g} i\, p_{x+y}(i) \tag{4.12}$$

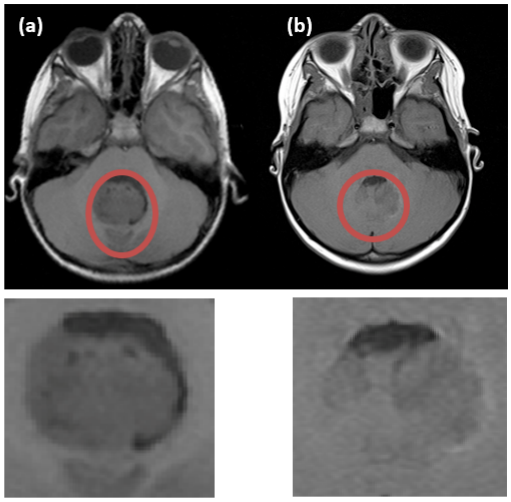
- *Sum Variance (SumVar)*: SumVar measures the variance of p_{x+y}

$$SumVar = \sum_{i=0}^{2N_g} (i - SumAvg)^2 p_{x+y}(i) \tag{4.13}$$

- *Sum Entropy (SumEnt)*: SumEnt measures the entropy of p_{x+y}

$$SumEnt = -\sum_{i=0}^{2N_g} i \log p_{x+y}(i) \tag{4.14}$$

Additional features also include *Difference Variance (DifVar)* and *Difference Entropy (DifEnt)*.



Contrast	6.28	36.82
Correlation	0.97	0.81
Inverse Difference Moment	0.43	0.24
Difference Variance	3.03	19.24

Figure 4.6: Two T1-weighted MR images of a (a) medulloblastoma and a (b) pilocytic astrocytoma. Tumour regions are marked in red. A close-up of each tumour site is shown beneath the MR images. The values of four GLCM features calculated from the tumour regions are shown below their corresponding images (GLCM direction: vertical; distance: 1 pixel). Original images were obtained from CCLG database [4].

To illustrate how GLCM features can quantify imaging patterns that could potentially aid the diagnosis of tumours, consider Figure 4.6. The figure shows two T1-weighted MR images of a MB and a PA. A GLCM was calculated for the tumour regions using a vertical direction of analysis ($\theta = 90^o$) and a distance of one pixel. The values obtained for four features *(contrast, correlation, inverse difference moment and difference variance)* are listed below their corresponding images. By inspecting the calculated feature values, one could see how the PA tumour site had considerably higher contrast and difference variance values than the MB site, suggesting that the former has a coarser, more heterogeneous texture.

(d) Grey-Level Run-Length Matrix

Shortly after GLCM was introduced as a means of quantifying textural patterns, Galloway [30] proposed the use of *grey-level run-length matrix* (GLRLM), a higher-order statistical technique. GLRLMs aim to capture information about the run[2] of a particular grey-level value or range of values, in a particular direction. Coarse textures are generally characterised by having short runs, whereas relatively longer runs populate fine textures. Similar to GLCMs, GLRLMs are commonly calculated for multiple combinations of directions (d) and distances (θ) of analyses (1 to 4 pixels along the $0^o, 45^o, 90^o and 135^o$ directions). The number of runs r with grey-level i, of run-length j, in a direction can be denoted by $R(\alpha) = [r'(i, j|\alpha)]$ [30]. Figure 4.7 shows a hypothetical image and its corresponding GLRLM.

[2] The number of pixels contained within the run is the run-length.

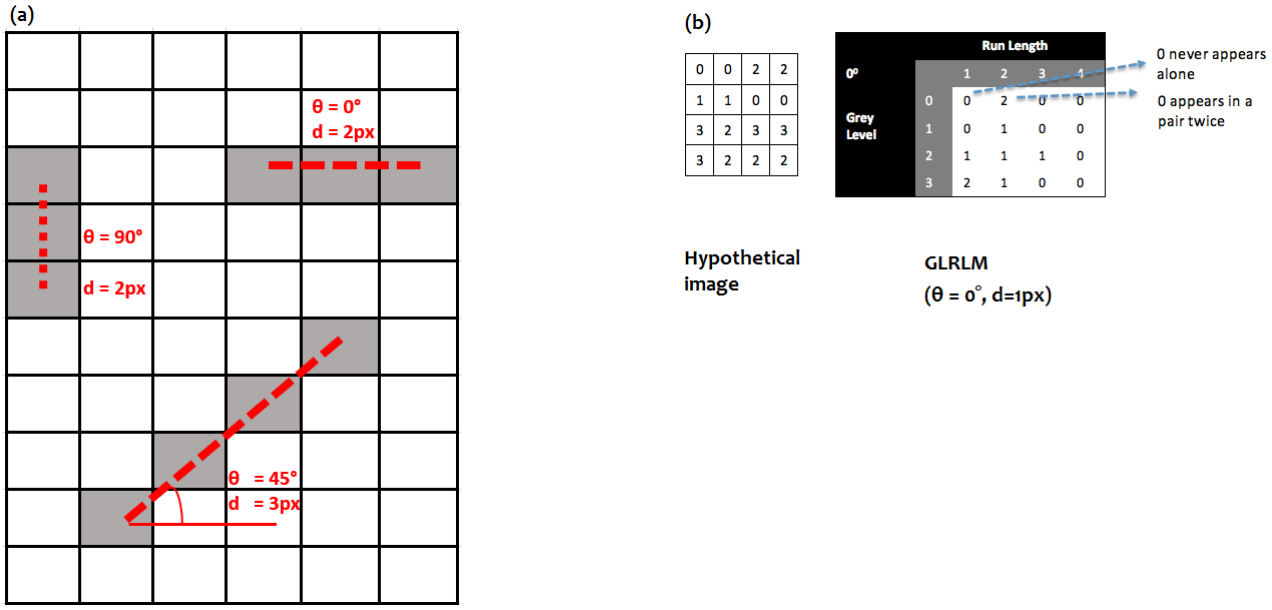


Figure 4.7: (a)Illustration of the pixel relationships considered by the Grey-Level Run-Length Matrix (GLRLM) technique.(b)A hypothetical image and its corresponding GLRLM, assuming a horizontal direction of analysis and a 1-pixel distance

.

Assuming that

- $p(i,j)$ is the number of times there is a run of length j having a grey-level i.
- N_g is the number of distinct grey-levels in the image ($i, j = 1, 2, ..., N_g$).
- N_r is the number of runs and P is the number of points in the image.
- Coefficient C is defined as $\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)$, i.e. the total number of runs in the image.

Then features extractable from a GLRLM can be computed as follows (Equations 4.15 to 4.19):

- *Short Run Emphasis (SRE)*: Measure of the proportions of runs that have short lengths. Expected to have large values in coarse textures.

$$SRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)/j^2}{C} \tag{4.15}$$

- *Long Run Emphasis (LRE)*: Measure of distributions of long runs. Assumes high values for smooth textures.

$$LRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)j^2}{C} \tag{4.16}$$

- *Grey-Level Non-Uniformity(GLNU)*: Measure that takes low values when runs are uniformly distributed along grey-levels.

$$GLNU = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_r} p(i,j)^2 \right)}{C} \tag{4.17}$$

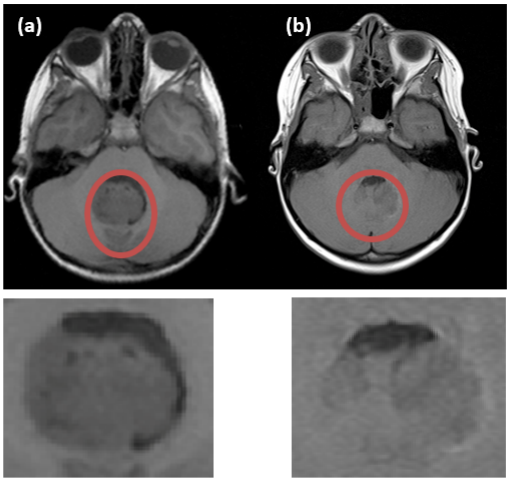
- *Run-Length Non-Uniformity (RLNU)*: Measure of the degree of non-uniformity

within run-lengths.

$$RLNU = \frac{\sum_{i=1}^{N_r} \left(\sum_{j=1}^{N_g} p(i,j)^2 \right)}{C} \tag{4.18}$$

- *Run Percentage(RP)*: This is the ratio of the total number of calculated runs to the total number of possible runs.

$$RLNU = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_r} p(i,j)^2 \right)}{P} \tag{4.19}$$



Grey-level non-uniformity	139	56
Run-length non-uniformity	3281	1113
Long run emphasis	2.06	1.36
Short run emphasis	0.86	0.93

Figure 4.8: Two T1-weighted MR images of a (a) medulloblastoma and a (b) pilocytic astrocytoma. Tumour regions are marked in red. A close-up of each tumour site is shown beneath the MR images. The values of four GLRLM features calculated from the tumour regions are shown below their corresponding images (GLRLM direction: horizontal; distance: 1 pixel). Original images were obtained from CCLG database [4].

Figure 4.8 shows two T1-weighted MR images of a medulloblastoma and a pilocytic astrocytoma (the same scans used in Fig 4.6). A GLRLM was calculated for the tumour regions using a vertical direction of analysis ($\theta = 90^o$) and a distance of 1 pixel. The values obtained for four features (grey-level non-uniformity, run-length non-uniformity, long run emphasis and short run emphasis) are listed below their corresponding images. The differences in the obtained feature values between the two tumours demonstrates how quantitative TA techniques can potentially provide decision support tools for tumour characterisation.

4.2.3 Three-dimensional Texture Analysis

Whilst most of the MRI TA experiments reported in the cancer literature focus on the analysis of textural features derived from a limited tumour area (a single 2D image slice) [20], there have been recent efforts to extend analysis to multiple MR image slices, as the processing of multi-slice volumetric features may offer additional information [33], [34], [35]. Intratumoural heterogeneity is likely to be greater in the whole tumour as compared to a limited region; hence, the use of the conventional 2D approach could dilute the diagnostic and prognostic value of TA [35]. In 2D TA, each particular voxel of interest has a maximum of 8 immediate neighbouring voxels that can be analysed in four independent directions $(0^o, 90^o, 45^o, 135^o)$. In 3D, each voxel of interest has up to 26 immediate neighbours, which increase the number of potential analysis directions to 13 [12]. This spatial relationship is illustrated in Figure 4.9, where the voxels of interest are visualised in red. Deciding which slice to include in the analysis is another limitation of conventional 2D TA.

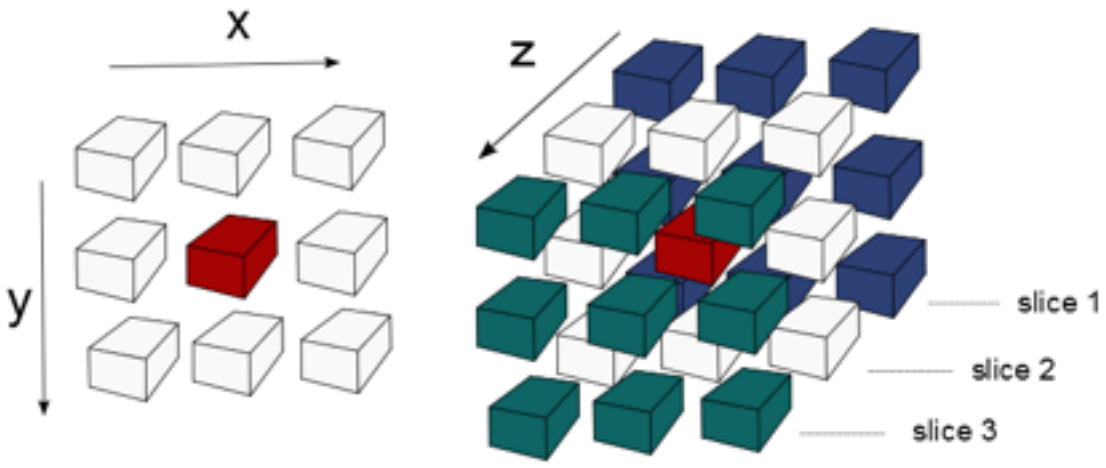


Figure 4.9: An illustration of the spatial relationship between voxels on a single two-dimensional image slice (left) and a three-dimensional multi-slice volume (right).

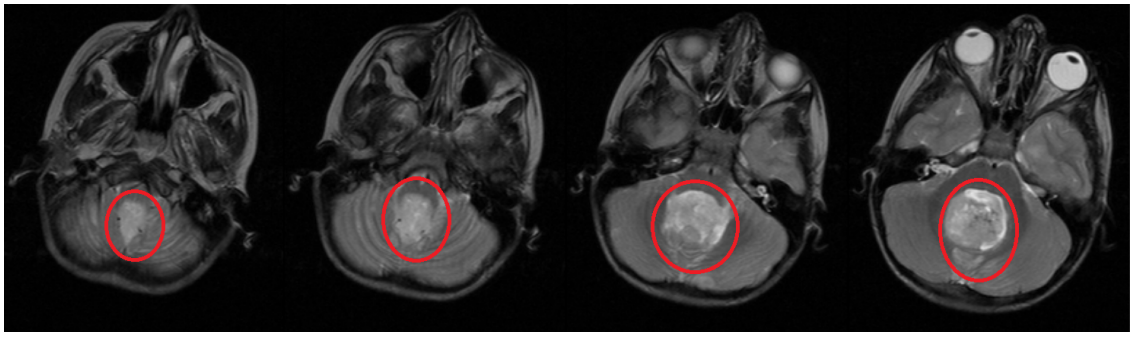


Figure 4.10: Multiple axial T2-weighted MR slices for one child diagnosed with medulloblastoma. Tumour regions are marked in red. Original images were obtained from CCLG database [4].

Figure 4.10 shows multiple T2-weighted axial slices for one child diagnosed with medulloblastoma (the images were obtained from the CCLG database). By inspecting the tumour regions, it becomes clear how appearance and texture can vary across multiple slices. Using only one MR slice might not be sufficient for building a reliable computational model, as capturing any heterogeneities present across the tumour volume would not be possible. In addition to this, 3D TA has the advantage of capturing inter-slice features that are completely ignored in the traditional 2D approach.

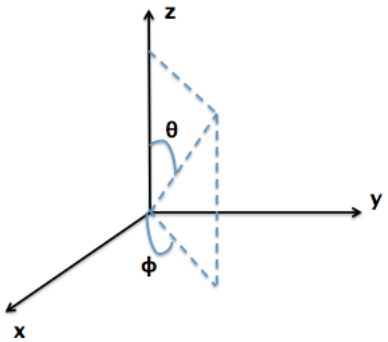


Figure 4.11: A figure illustrating θ and ϕ, which are used to spatially characterise directions of analysis in 3D GLCMs [111].

Grey-Level Co-Occurence Matrices (GLCMs), which were described in section 4.2.2, can be extended to capture the spacial dependencies of grey-level values

Table 4.1: Offsets describing 13 possible directions of analysis when computing 3D GLCMs and GLRLMs. d is the chosen distance of analysis, in number of pixels.

	Offset	Degree Direction (θ and ϕ)
1	0,-d,0	0^o,0^o
2	d,-d,0	45^o,0^o
3	d,0,0	90^o,0^o
4	d,d,0	135^o,0^o
5	0,-d,d	0^o, 45^o
6	0,0,d	0^o,90^o
7	0,d,d	0^o,135^o
8	d,0,d	90^o,45^o
9	-d,0,d	90^o, 135^o
10	d,-d,d	45^o,45^o
11	-d,d,d	45^o, 135^o
12	d,d,d	135^o,45^o
13	-d,-d,d	135^o,135^o

across multiple slices. Like the conventional 2D GLCMs, this matrix also acts as an accumulator, where *p[i,j]* counts the number of pixel pairs having intensities i and j. The presence of a third dimension, however, means that GLCM is no longer computed using the conventional 4 directions of analysis. In the volumetric case, there are up to 13 unique directions of analysis. To spatially characterise them, two angles are used: θ and ϕ [111], as depicted in Figure 4.11. The 13 possible directions can be described using unique offsets, as summarised in Table 4.1. Directions 1-4 are those computed using the conventional 2D GLCM approach.

It is worth noting that, using the same principles, Grey-Level Run-Length Matrices (GLRLMs) can be extended to capture multi-slice, volumetric pixel patterns. The same 13 directions summarised in Table 4.1 can be used to construct the matrix. When calculating 3D GLCM and GLRLM, each slides does not need to be processed individually, but all slices are processed at once, producing only one matrix for all consecutive slices forming the 3D images [112].

4.2.4 Practical Limitations of MRI Texture Analysis

It is clear from the discussion above that 3D TA would ideally require minimal MR image slice gaps, for volumetric features to capture characteristics of maximal value. This introduces a fundamental limitation within the context of this thesis, where clinical data used is retrospective multi-slice MR scans that had been acquired using conventional Spin-Echo sequences (*i.e. the images are not true 3D*; Chapters 5 -7).

Additionally, the use of slices for image acquisiton means that each slice summarises potentially many different elements of the underlying pathological structure over its width [21]. This leads to the interesting question of whether selecting thinner slices during acquisition might lead to imaging data that can build more robust TA predictive models[3]. In this regard, it is likely that between-plane textural attributes that will be calculated via 3D TA are going to be lacking in robustness; and if 3D TA is able to yield better classification then it will likely arise due to the inclusion of more representative data than those obtained via a single slice.

Another related limitation is the issue of resolution dependency. In practical clinical settings, an image with dimensions 256 by 256 pixels is commonly used due to reasonable collection time (i.e. 256 times TR, approximately 6-14 minutes). If image dimensions are increased to 512 by 512 pixels or 1024 by 1024 pixels, the data collection time will become considerably longer, possibly reaching 28 minutes [21]. With a field of view 230-240 mm, which is usually used in whole body scanners, pixel size will be around 1mm, 0.5mm or 0.25mm respectively. Thus, subtle variations in imaging characteristics are likely to arise between images of various pixel sizes, introducing a practical limitation for implementing TA in clinical settings, as feature meaning would not be consistent across various resolutions.

[3]The use of thinner slices, however, will result in a worsened signal-to-noise ratio, thus concealing the true texture [21].

4.3 Review of the Current State-of-the-Art

The material presented in this section provides an extensive review that covers TA work available in the adult and childhood cancer literatures.

4.3.1 Applications of TA in Diagnostic Classification of Paediatric Brain Tumours

Below is an up-to-date review that summarises studies that looked into using MRI TA for diagnostic classification of childhood brain tumours, to the best of my knowledge at the time of writing the thesis. The reviewed articles are summarised in Table 4.2 at the end of this section.

Rodriguez Guiterrez et al [36] recently studied the performance of a support vector machine (SVM) classifier that was trained with 2D textural features in order to classify paediatric posterior fossa tumours. Features consisted of conventional histogram statistics as well as second order GLCM features, which extracted from T1, T2 and diffusion-weighted MR images. Besides aiming to classify tumours into one of three main classes (medulloblastoma, pilocytic astrocytoma and ependymoma), the study also looked into the classification of tumour sub-types. A cohort of 40 patients was used. Encouraging tumour type classification rates that ranged between 71% and 84% using T1 and T2-weighted images were obtained. An improved tumour type classification performance was obtained when diffusion data was used, yielding up to 91% accuracy. The study reported a classification accuracy of up to 89% with regards to the tumour-subtype classifier. Nevertheless, one should note that the tumour subtypes included in the study were highly imbalanced (e.g. 14 classic vs. 3 anaplastic medulloblastoma), yet no statistical measures (e.g. minority over-sampling) were implemented to mitigate this. The

reported results are therefore unlikely to resemble the real-world tumour subtype classification performance. In addition to this, the study did not report on other evaluation metrics, particularly *sensitivity, specificity or area under the curve*, which can give a more realistic insight of the obtained results.

A similar study was conducted by Orphanidou-Vlachou et al [37], where an artificial neural network (ANN) was trained with 2D textural features in order to classify paediatric posterior fossa tumours. The textural features used were based on histogram statistics, absolute gradient, GLCM, GLRLM, autoregressive model and Haar wavelets. Both T1 and T2-weighted MR images were included in the analysis. In order to reduce feature dimensionality, principal component analysis (PCA) was performed prior to evaluating classifier performance. Since a small cohort of 40 patients was used, model validation was performed using leave-one-out cross validation (LOOCV); and ten-fold cross validation (CV) was also carried out in order to provide additional reassurance. On both LOOCV and ten-fold CV, ANN was able to achieve 90% classification accuracy. In order to obtain a radiological benchmark, the authors reviewed conventional radiological reports of the cases included in the study. The review showed that the correct diagnosis was specified in only 45% of the reports. Even in cases in which the correct diagnosis was specified, 22% of them had an alternative proposed. Nonetheless, the review also showed that incorrect diagnosis was only specified in 5% of the cases, which suggests a large degree of diagnostic uncertainty in conventional radiological reporting. It is worth noting that it is not the job of the radiologist to diagnose the tumours - another likely explaination for limited number of cases where a correct diagnosis was specified. In terms of limitations of the reported model, carrying out data-driven pre-processing (PCA) on the entire dataset prior to, and not within, the cross-validation loop means that the reported results are perhaps

not indicative of the **overall** process. To get a more representative estimate of the final model, PCA would ideally need to be carried out on the surrogate training-set, separately for each corresponding model. Nevertheless, the overall findings of the study suggest that the use of a non-invasive diagnostic aid like TA can potentially improve radiological diagnostic confidence and performance.

In an effort to apply TA on multi-modal and multicentric datasets, Tantisatirapong et al [38] analysed T1, T2, FLAIR, diffusion-weighted and diffusion-tensor images obtained from four different hospitals. Two types of tumours were considered in this study: medulloblastoma and pilocytic astrocytoma. Unlike the work by Rodriguez Guiterrez et al [36] and Orphanidou-Vlachou et al [37], the included cases were not restricted only to tumours of the posterior fossa. Although the total number of cases was 50 (25 MB and 25 PA), complete datasets were only available for conventional T2-weighted images as not all patients had images available from all modalities. Similar to the work reported in [36], an SVM was designed to carry out tumour type classification. Supervised feature selection was carried out using the sure independence screening technique, followed by LOOCV, which was used for model performance evaluation. Classification results showed that TA on diffusion data yielded higher performance compared to when conventional MRI data was used. For instance, a classification accuracy of up to 97% was obtained when diffusion-weighted images were used, compared to 77% with T2-weighted images. However, the heterogeneity of the dataset meant that only 25 cases were available for diffusion-weighted images, compared to 50 T2-weighted images. Such heterogeneity impedes the interpretation of the study's findings and makes it difficult to agree that TA on certain modalities is superior to others. It is also worth noting that the feature selection technique used (sure independence screening) requires knowledge of class labels to provide a rank of feature importance. Thus, including

it before, and not within, the cross-validation loop is likely to have introduced an element of over-optimistic bias to the evaluated model. Nonetheless, the study's findings generally support the use of MRI TA as a non-invasive aid for diagnosing childhood tumours. Additionally, the multicentric nature of the data used suggests that TA is possibly a scalable technique that allows transfer of results across centres.

Table 4.2: TA articles available in the literature that look into classifying childhood brain tumours from MR images.

Author	Modalities	Cohort Size	Multicentric?	2D/3D	Methods	Accuracy	Study Limitations
Rodriguez Guiterrez et al (2014)	T1, T2 and Diffusion-weighted.	40	No	2D	SVM to classify MB, PA and EP.	71% to 84% on T1 and T2.	Only looked into posterior fossa tumours.
Orphanidou Vlachou et al (2014)	T1 and T2-weighted.	40	No	2D	ANN and LDA to classify MB, PA and EP.	MB (94%), PA (81%) and EP (63%)	Only looked into posterior fossa tumours.
Tantisatirapong et al (2013)	T1, T2, diffusion-weighted, FLAIR and DTI.	50	Yes	2D	SVM to classify MB and PA.	77% on T2.	Large amounts of missing data.

4.3.2 Applications of TA in Diagnostic Classification of Adult Brain Tumours

This section reviews articles available in the literature that look into the classification of adult brain tumours using TA of MR images. At the end of the section, a summary of the reviewed articles is provided in Table 4.3.

The problem of brain tumour classification has been of interest in the adult literature since the nineties. Early work reported by Lerski et al [39] did not look into classifying tumour types, but rather into the more generic problem of classifying brain tissue types (e.g. oedema, white-matter, grey-matter). In this prospective study, 12 patients with brain tumours were examined. A total of 78 ROIs were defined from T1 and T2 parameter images of the selected slices: 8 tumour, 11 oedema, 12 liquor, 24 white-matter and 23 grey-matter. Using a hierarchical decision tree, a classification system that uses textural features was

developed in order to discriminate between the tissue types. The study reported classification accuracies that ranged between 74% and 100%, demonstrating the existing potential of using TA of MR imaging as a tool for brain tissue characterisation. However, one may question the statistical reliability of the reported results, since using 78 ROIs to represent data from 12 patients might have lead to an overoptimistic classification performance.

Ten years later, Mahmoud-Ghoneim et al [34] looked into the problem of classifying different tumour regions (e.g. edema and necrosis) using TA. The study was a preliminary evaluation that only considered gliomas. Using a linear discriminant analysis (LDA) classifier, 2D and 3D GLCMs were compared in characterising between necrosis, solid tumour and edema from T1-weighted scans. This was one of the first articles to propose the use of 3D TA of MRI in brain cancer, by extending the analysis to include the tumour volume, rather than a single 2D slice. The primary conclusion of the study was that the use of 3D GLCM outperformed the conventional 2D approach. Nevertheless, since only 7 cases were included, practical applications of the reported findings are perhaps not immediate.

An interesting approach was followed in the study by Glotsos et al [40], where TA was used for tumour grade classification using digitised biopsy images, rather than MR or CT scans. A large cohort of 140 astrocytoma biopsies was included in the study, and an SVM classifier was used to identify the grade of the images (Grades II, III or IV). Although long-term clinical benefits of such tool might not include reduction of surgery, as the analysis was done on biopsy images, the proposed methodology might be a great aid that can be used in parallel with conventional histopathological grading to support the regular diagnostic procedure.

Georgiadis et al reported a number of studies that looked into classifying adult brain tumours from MR images using TA methods. The first study [41] evaluated

the performance of a software system designed to discriminate between metastatic and primary brain tumours (gliomas and meningiomas) using textural features of contrast-enhanced T1-weighted images. 67 patients were included in the study. Using a least square feature transformation probabilistic neural network (LSFT-PNN), a two-level classification system was designed. The system worked by separating primary and secondary brain tumours at the first level, followed by further classification between gliomas and meningiomas if the tumour was classified as primary. Using an external cross-validation scheme, the system was able to yield classification rates that ranged between 74% and 89%. Demonstrating the feasibility of using TA to discriminate between primary and metastatic tumours is of clinical value because metastatic tumours require specific treatment protocols (e.g. radiation therapy), whereas primary tumours may also require surgical intervention [89], [90].

Georgiadis et al [33] carried out a second TA study on brain MR images of adult patients, using histogram statistics, GLCM and GLRLM. The included cohort consisted of 67 post-contrast, T1-weighted scans. An SVM-based classification system, which was trained with textural features, was able to discriminate metastatic, malignant and benign tumours with 77%, 89% and 93% classification accuracies respectively. Note that a sub-objective of the study was to compare the performance of conventional 2D TA to 3D TA, which involves extending the analysis to include the full tumour volume. The study showed that 3D TA was able to outperform the conventional 2D approach when discriminating primary from metastatic tumours (94% vs. 89% accuracy).

In an effort to fuse information obtained from different MR modalities for the characterisation of brain cancer, the third study carried out by Georgiadis et al [42] combined 3D textural features of T1-weighted images with MR spectroscopy. The

textural features were based on histogram statistics, GLCM and GLRLM. From the spectroscopic data, three metabolites were considered: choline (Ch), N-acetyl aspartate (NAA) and creatine (Cr), and the following ratios were used as features: Cho/NAA, Cho/Cr and NAA/Cr. Using a cohort of 40 patients, the use of an SVM classifier trained with the combined features was able to yield a classification accuracy of 91% when discriminating between meningiomas and metastatic brain tumours.

Another interesting study that used multimodal imaging information was carried out by Zacharaki et al [25], with the aim of designing a classification system that could distinguish different brain tumour types, and also grade gliomas. 98 patients were included in the analysis. In terms of MR modalities, the following imaging types were used: T1, contrast-enhanced T1, T2, FLAIR and rCBV maps. The classification system was based on a large number of features (161), which included age, tumour shape characteristics, image intensity characteristics within some of the ROIs, as well as features extracted using TA. Three classification algorithms were considered: SVM, LDA and k-Nearest Neighbours (KNN). The best performance was achieved using SVM in the binary classification scenario, with accuracies that ranged between 72% and 96%.

Table 4.3: A summary of TA articles available in the literature that look into classifying adult brain tumours from MR images.

Author	Modalities	Cohort Size	Multicentre	2D/3D	Methods	Accuracy	Study Limitations
Lerki et al (1993)	T1, T2 parameter images.	12	No	2D	Hierarichal Decision Tree to classify different brain tissuires.	74% to 100%.	Used 78 ROIs to represent 12 patients.
Mahmoud-Ghoneim et al (2003)	T1-weighted.	7	No	3D	LDA to classify different tumour regions.	57% to 86%.	Only 7 cases were included.
Glostos et al (2005)	Digitised biopsy images.	140	No	2D	SVM to classify tumour grade.	84% to 88%.	Using biopsy images requires invasive surgery.
Georgiadis et al (2008)	Contrast-enhanced T1-weighted.	67	No	2D	PNN to classify gliomas, meningiomas and metastatic tumours.	74% to 89%	
Georgiadis et al (2009)	Contrast-enhanced T1-weighted.	67	No	3D	SVM to classify benign, malignant,and metastatic tumours.	77% to 93%	Improvements by 3D TA were not tested for statistical significance.
Georgiadis et al (2011)	T1-weighted and MRS.	40	No	3D	SVM to classify meningioma from metastatic tumours.	91%	Improvements by new system were not tested for statistical significance.
Zacharaki et al (2009)	T1, contrast-enhanced T1, T2, FLAIR and rCBV maps.	98	No	2D	SVM, LDA and KNN to distinguish between different tumours and grade gliomas.	72% to 96%.	

4.3.3 Other Diagnostic Applications of MRI TA

Although the focus of this thesis is on brain tumour characterisation, it is worth noting that TA of MRI was shown to be of value in other clinical applications. For example, Chen et al [26] investigated the efficacy of 3D GLCM for the characterisation of breast MR lesions. Using T1-weighted data of 121 lesions, it was shown that 3D textural features can significantly outperform those based on 2D analysis.

In another breast cancer classification study, Holli et al [43] successfully used TA to distinguish between healthy and cancerous breast tissues from contrast-enhanced T1-weighted images. Additionally, different histological types of breast cancer (lobular and ductal) were successfully discriminated using textural features.

TA of multimodal MR images (T1, T2 and diffusion-weighted) was also shown effective in discriminating between healthy and cancerous prostatic tissues [44]. With regards to applications in other pathological conditions, the potential value of TA was demonstrated in Alzheimers disease [45], multiple sclerosis [46] and epilepsy [47].

4.3.4 TA for Estimating Survival Prognosis

To the best of my knowledge at the time of writing, there has been no published work on investigating brain tumour survival predictors based on image analysis of conventional MRI, such as T1 and T2-weighted scans. Such scans are routinely acquired when a patient is presented with a suspected brain tumour, and their reported success in diagnostic TA applications suggests a possibility that valuable but complex prognostic patterns may exist undiscovered in the data.

However, TA was used successfully for estimating survival prognosis in a different problem domain, namely lung cancer, as per Ganeshan et al [48]. The

study reported the use of textural features that measure tumour heterogeneity as a way of quantifying computed tomography (CT) scans pixel distributions. These features were based on the Laplacian of Gaussian filters technique, and ranged between fineness (homogeneity) and roughness (heterogeneity) attributes. The authors argued that heterogeneity is a good measure for estimating prognosis because it is a well-recognised feature of malignancy that is associated with adverse tumour biology. For instance, heterogeneity of the tumour blood supply is associated with lack of sufficient oxygen supply and genomic instability. Based on this, the authors hypothesised that biological heterogeneity can be reflected on CT scans and can consequently be measured using textural features. The results of the study supported the primary hypothesis, with heterogeneous tumours demonstrating significantly poor survival patterns.

TA of CT scans was also successfully employed as part of other studies that investigated potential markers for oesophageal cancer [49] and colorectal cancer [51]. Both studies employed a similar methodology to that reported in [35], where the extracted features were based on the Laplacian of Gaussian filter technique and ranged between fineness (homogeneity) and roughness (heterogeneity) features. Such results support the use of TA as a means of capturing imaging patterns that have the potential to predict survival prognosis, and encourage its adoption in the paediatric neuro-oncology domain.

4.4 Summary

This chapter provided a review of TA methods and studies available in the MRI literature. The first section reviewed commonly used statistical methods, namely histogram analysis, absolute gradient, GLCM and GLRLM. The features extractable from these techniques were also explained, together with mathematical formulae

that can compute them. 3D (multi-slice) texture analysis, which aims to maximise information extractable from MR images, was also explained.

The second section provided a critical review of MRI texture analysis articles available in the literature. The covered work included studies that looked into diagnostic classification of paediatric and adult brain tumours from MR images. Three studies reported success in using conventional 2D TA to diagnose different childhood brain tumours. The adult literature reported a number of studies that showed improved performance when 3D TA was used. There exists, therefore, considerable motivation for further research into maximising the value of TA as a predictive biomarker in paediatric oncology using the 3D approach. Whether the value of TA can be maximised in paediatric settings is as yet largely untested, but will be needed to translate the role of TA in clinical practice.

A key finding of the literature review was that none of the studies available in the paediatric and adult MRI literature looked into analysing the survival of patients diagnosed with brain tumours. Nevertheless, TA was successfully applied on CT scans to estimate the survival in different problem domains, namely lung, oesophageal and colorectal cancer.

Chapter 5

A Single Centre Study on 3D TA

Some aspects of the work presented here were published in [**P01**] and [**P03**]. Publication details can be found on Page *xx*.

5.1 Introduction

This chapter presents a thorough investigation into the diagnostic efficay of MRI TA in paediatric settings.

As concluded from the literature review, to the best of the author's knowledge at the time of writing, the current state-of-the-art in the paediatric literature is the conventional 2D TA approach, which can potentially dilute the diagnostic value of textural attributes. None of the 3D MRI TA work reported in the current literature was carried out on paediatric brain tumours. In this regard, the primary aim of the study presented here was to investigate the effectiveness of carrying out 3D TA on T1 and T2-weighted MR images for classifying paediatric brain tumours. This was done by testing the performance of six different supervised classification algorithms trained with 3D textural information in differentiating between medulloblastoma, pilocytic astrocytoma and ependymoma - the most common types of brain tumours occuring in childhood. It was hypothesised that carrying out the analysis in three dimensions would yield more discriminative information about the tumours than the traditional 2D approach.

In addition to this and in light of the No Free Lunch theorem, which states that there is no universal classification model that works best for every problem domain, it is sensible to explore different classifiers in order to identify the ones most likely to generalise well with our data. Therefore, a sub-objective of this study was to compare the performance of several models that represent typical implementations of supervised classifiers. Studying different classifiers is a fair way of determining whether any improvements due to use of 3D textural features

would be consistent throughout different models.

5.2 Materials and Methods

5.2.1 Cohort Details and Image Acquisition

The datasets used in this study were obtained from a secure e-repository provided by the Childrens Cancer and Leukaemia Group (CCLG) Functional Imaging Group [4]. The dataset consisted of pre-contrast T1 and T2-weighted MR images of 48 children (31 male, 17 female) with untreated brain tumours, of which 21 were medulloblastomas (MB), 20 were pilocytic astrocytomas (PA) and 7 were ependymomas (EP). In terms of tumour characteristics, all MBs were located in the vermis of the posterior fossa, and ranged between 20 and 46 mm in size. Four PAs were located in the middle fossa (28-69 mm), and the rest were in various locations of the posterior fossa (23-78mm). Two EPs were located in the middle fossa (42 and 48mm), while the rest were in the posterior fossa (32-45mm).

Image acquisition was carried out at a single centre[1], using a spin-echo sequence on a GE Signa 1.5 T scanner (GE Healthcare, Little Chalfont, UK) and a Siemens Symphony 1.5 T scanner (Siemens Healthcare, Erlangen, Germany). For T1-weighted images, echo time was 8.4-22 ms, repetition time was 360-819 ms, slice thickness was 4-5 mm, slice gap was 0.8-1.5 mm and image resolution was 1.391-2.560 pixels/mm. For T2-weighted images, echo time was 77-105 ms, repetition time was 3940-7840 ms, slice thickness was 3-5 mm, slice gap was 0.6-1.5 mm and image resolution was 1.948-2.560 pixels/mm.

Approval for the study was obtained from the research ethics committee, and informed consent was taken from guardians. All data had been anonymised before

[1]Birmingham Children's Hospital.

uploading to the database. In order to obtain diagnoses in accordance with the WHO classification, tumour samples were taken from all patients and underwent histopathological examinations.

5.2.2 Image Pre-processing

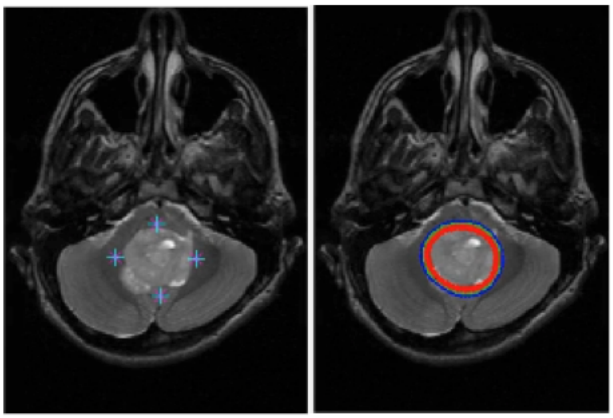


Figure 5.1: A figure showing semi-automatic segmentation of a tumour region of interest using the Snake GVF algorithm. Initial seeding points are shown on the left and final contour movements are shown on the right. Original MR images were obtained from the CCLG database [4].

Axial slices were manually chosen from each dataset using RadiAnt DICOM viewer [61]. Semi-automatic segmentation was performed on MATLAB (MathWorks, Massachusetts) using the snake gradient vector flow (Snake GVF) technique (Figure 5.1), as proposed by Xu and Prince [72], in order to extract the regions of interest (ROIs) in which the tumour was present. The *Snake: Active Contour* library was used for this [94]. The medical imaging literature contains a plethora of work conducted on studying the effectiveness of automatic and semi-automatic segmentation algorithms. Whilst the implementation of an effective segmentation method is crucial for capturing tumour information, the details of which approach yields optimal segmentation is not of immediate interest within the context of this thesis. The Snake GVF technique works by relying on man-

ually defined seeding points, which are initially outlined by the user [72] . The segmentation boundary is then constructed by calculating an edge map of the input image and progressing the contour towards a so-called force balance condition, where an internal force that prevents contour stretching is balanced with an external force that pulls the snake towards the desired contour. After segmentation, the ROIs were checked visually to ensure that the segmentation technique worked sufficiently well.

The ROIs were imported to MaZda texture analysis software, which was developed by Materka and co-workers as a part of the European COST B11 and the following COST B21 programs [31]. The choice of the software was quite deliberate since it has been extensively used in the MRI TA literature [27],[37],[43],[95],[96], [97],[98]. To mitigate the variations in parameter settings used while scanning different patients, the grey-level values within the identified ROIs were normalised. In MaZda, normalisation is a two-step process that requires:

- Grey-level range selection.
- Image quantisation, by re-sampling to a certain number of bits per pixel.

The first step (range selection) was carried out using the limitation of dynamics to $\mu + / - 3\delta$ (where μ is the ROIs mean grey-level value and δ is the standard deviation), which was shown by Collewet et al to achieve reliable results on MRI texture classification [63]. This method works by computing the $\mu - 3\delta$ and $\mu + 3\delta$ values from each ROIs histogram, and excluding any values that lie outside that range. This step does not stretch out or compress the histogram, it simply decides on which grey-levels to include and exclude from the range. The second step involves quantising the resulting grey-level range between 1 to 2^k , where k is the number of bits per pixel. For instance, if our original range is between 1 and 1024,

but we choose to use 8 bits per pixel, the dynamic range would be quantised to the range 1 to 256. In this study, 6 bits were chosen for quantisation.

5.2.3 Textural Features Extraction

MaZda [31] was used to perform both 2D and 3D TA on the normalised ROIs segmented from T1 and T2-weighted images. In traditional 2D analysis, one T1 and T2-weighted ROI was used for each patient to calculate intra-slice metrics and their corresponding features, i.e. the features represent only the ROI of the chosen slice. Calculations were carried out on the ROI from the slice that contains the largest tumour area. For 3D analysis, multiple adjacent T1 and T2-weighted ROIs were used to calculate metrics that hold intra-slice and inter-slice pixel relationships. The TA methods, together with the extracted features used, are listed in Table 5.1.

Table 5.1: A table summarising the TA methods used and their corresponding features.

TA Methods	Calculated Features
Histogram statistics	Mean, variance, skewness, kurtosis, minimum, maximum and percentiles (1%, 10%, 50%, 90% and 99%).
Absolute gradient statistics	Absolute,gradient mean, variance, skewness and kurtosis.
Grey-level co-occurrence matrix (GLCM)	Ang. Second Moment, inverse difference moment, contrast, correlation, entropy, sum of squares (variance), sum average, sum variance, difference variance and difference entropy.
Grey-level run-length matrix (GLRLM)	Short run emphasis, long run emphasis, grey-level non-uniformity, run length non-uniformity and run percentages.

5.2.4 Feature Selection and Analysis

The features that were computed using the above techniques were aggregated for analysis, giving us two feature sets; one that holds 2D T1 and T2 textural features, and a second one that holds 3D T1 and T2 textural features. A breakdown of the number of features extractable from each technique is shown in Table 5.2. The feature sets were imported to Orange, the python-based machine-learning library

(version 2.7) [60], which was used to analyse and compare the two sets separately.

Table 5.2: A table showing a breakdown of the number of textural features for each dataset.

	2D T1	2D T2	3D T1	3D T2
GLCM	191	191	240	240
GLRLM	19	19	24	24
Histogram	12	12	13	13
Absolute Gradient	5	5	6	6

Feature Selection

As discussed in Chapter 3, testing all possible combinations from all techniques, modalities and datasets would give a very large number of features (454 for 2D and 566 for 3D). Entropy-MDL discretisation was used to partition our textural features to a discrete number of intervals. The discretised feature sub-set holds only the features that the algorithm deduced to be the most relevant and discriminative; since a features entropy can be used as a measure of its discriminative power. Given the supervised nature of this technique, where class-label information is used to determine discretisation cut-off values, this method should ideally be implemented within cross-validation loops; otherwise, an element of over-optimisim would be introduced to the designed models.

Supervised Learning

The 2D and 3D sub-sets holding the selected features were separately analysed using the six supervised machine-learning classifiers listed below, which were introduced in Chapter 3. Python's Orange library [60] was used to implement the classifiers.

- Naive Bayes (NB): Prior class probabilities based on: *Relative frequency*.
- K-Nearest Neighbour (kNN): Neighbours: *5*, Distance metric: *Euclidean*.
- Classification Tree (CTr): Attribute selection based on: *Information gain*.
- Support Vector Machines (SVM): SVM Type: *C-SVM*, Kernel: *RBF*.
- Artificial Neural Network (ANN): Number of hidden layers: *1*.
- Logistic Regression (Logreg).

Model Validation

Leave-one-out cross-validation (LOOCV) technique was used in order to evaluate the classification performance. Using LOOCV, learning sets were created by taking all the samples but one, which was used as the test set. This process was looped 48 times, in order to cover the 48 possible ways of obtaining such a partition in our dataset, and the results were then averaged. LOOCV was chosen because the training sets always contain only one element less than the full feature set, and hence the obtained predictive performance has, in theory, a good potential of closely reflecting the real performance. Classification accuracy, sensitivity, specificity and area under the receiver-operator characteristics curve were measured by calculating true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). In addition to using LOOCV, *stratified 10-fold cross-validation* (CV) was also performed for model validation in order to provide additional reassurance of the classifiers reliability. The stratified approach was chosen instead of conventional 10-fold CV in order to make sure that all three tumour classes are represented in the validation folds.

Examining an Alternative Dimensionality Reduction Approach (PCA)

It may be argued that an alternative dimensionality reduction technique, namely principal component analysis (PCA), needs to be considered. This method is arguably more intuitive than the feature selection technique we implemented (Entropy MDL) and has been extensively used in the literature. The use of PCA, within a supervised MRI TA investigation, to reduce the dimensionality of a 2D textural feature set was recently reported by Orphanidou-Vlachou et al [37]. The underlying principle behind PCA is based on building a new set of features (principal components) in a way that maximises the variance of the original feature-set and brings down its complexity. A PCA-based pipeline was therefore built and tested on Orange, for classifying the same datasets and using the same classifiers described above (Naive Bayes, Classification Tree, kNN, SVM, ANN and Logistic Regression). For the purpose of analysis, the principal components (PCs) covering 98% of the variance were chosen. In keeping with the methodology used throughout this experiment, PCA-based model validation was carried out using the LOOCV approach.

5.2.5 Statistical Analysis

Statistical Comparison between 3D and 2D Classification Results

Once the classifiers were trained and evaluated, the next logical step was to apply a statistical test, such as McNemar's test [64], in order to compare classification performance obtained with 3D and 2D TA. For each of the six classifiers, McNemar's test was carried out to test whether any improvements obtained with 3D TA were significantly different to the 2D results. The classification results used in the test were those obtained from LOOCV. The first step to apply McNemar's test is to construct a contingency table as shown in Table 5.6. The table's entries provide a summary of the number of agreements and disagreements between 3D and 2D-trained classifiers. For each of the six contingency tables that we constructed, we computed their corresponding chi-square value and tested the result against the theoretical chi-square with one degree of freedom. A power analysis calculation showed that 38 samples would be needed to achieve a high power of 0.90 for a two-tailed paired t-test at a significance level of 0.05. In doing this, prior information on classification accuracies was obtained from the adult brain cancer literature that compared 3D and 2D TA [33], [34], as the current study is the first to investigate this in childhood.

Statistical Comparison between Individual Classifiers

Besides maximising the diagnostic value of TA using 3D features, a sub-objective of this study was to determine whether the choice of classification algorithm has substantial influence over the results. To this end, pairwise comparisons using McNemar's test were carried out to establish whether the performance of different classifiers is significantly different. The test was carried out on the LOOCV results.

Testing for Over-fitting using Bootstrapping

It may be argued that the use of a cohort of 48 subjects is rather small given the number of textural features used, which can potentially lead to fitting instability or over-fitting. We therefore calculated confidence intervals of classification accuracies by bootstrapping the subjects in the sampling, thus addressing any concerns with regards to the use of a small cohort. The bootstrapping process was carried out by sampling the subjects with replacement, followed by equally splitting the sampled distribution into a training and a testing set. Upon applying feature selection and the learning algorithms, the obtained classification accuracy was then recorded for each tumour type. This was repeated 1,000 times in order to infer a realistic distribution of classification accuracies. The lower and upper bounds of the confidence intervals were chosen to be the accuracies at positions 2.5% and 97.5% of the sorted accuracies list, respectively. It remains noteworthy, however, that any bias introduced by carrying out supervised feature selection outside the cross-validation loop will not be corrected by bootstrapping.

5.2.6 Obtaining a Radiological Review Benchmark

For the cases included in the study, a review of corresponding radiological reports was carried out in order to get an insight into the reliability of relying solely on MR images for obtaining diagnoses. The reports were those used clinically. All of them were made by five consultants who worked in Birmingham Children's Hospital.

5.3 Results

5.3.1 Top ranked 2D features

108 out of the available 454 features were selected by entropy-MDL from the 2D dataset, as summarised in Table 5.3. It is worth noting that since GLCM features were computed via different combinations of directions and distances, each GLCM feature is represented using a unique 2-digit offset, which could be summarised as follows: $(0^0 = 0, D); (45^0 = D, -D); (90^0 = D, 0); (135^0 = D, D)$, where D represents the distance of analysis. For example, an offset [0, 1] would be an analysis of 1 pixel distance in the horizontal direction. Features' acronyms were introduced in Chapter 4.

The selected T1-weighted sub-set included variations of only three GLCM-based features: entropy, sum entropy and angular second moment. Selected T2-weighted features, however, included a wider variety of measures based on GLCM (such as contrast, correlation, inverse difference moment and difference variance). Additionally, a number of T2-weighted GLRLM and histogram based features were recognised as important.

5.3.2 Top ranked 3D features

122 out of the available 566 features were selected by entropy-MDL from the 3D dataset, as summarised in Table 5.4. Note that 3D GLCM features are represented using a three-digit offset rather than a two-digit one, where the third digit identifies whether the analysis took place along the z-axis (between slices) [64]. Both T1 and T2-weighted sub-sets included variations of features based on GLCM (such as entropy, sum variance and sum average), GLRLM and histogram statistics.

By inspecting tables 5.3 and 5.4 one can see that there was an element of

directional sensitivity when choosing important features. For instance, the T2-weighted 2D sub-set included angular second moment calculated in the directions 0, 2 and 2, 2, but not 2,0. It may be argued that feature selection should not demonstrate such sensitivities for GLCM and GLRLM features measured across different combinations of pixel distances and directions. We have therefore carried out Pearson Correlation on a number of GLCM features to explore whether the same features measured across different directions capture similar patterns. The findings are plotted on distance maps, which can be found in Figure 5.2. The findings suggest a very strong positive correlation between most of the tested features' variations across different distances, indicating that they capture strongly correlated patterns. There are, however, small variations between these features and so not all such highly correlated features will be selected, depending on the cut-off, as detailed in the discussion section.

Table 5.3: A table showing a summary of the T1 and T2-weighted 2D features chosen by entropy-MDL during the feature selection stage. Each GLCM feature is represented using a unique 2-digit offset as follows: $(0^0 = 0, D); (45^0 = D, -D); (90^0 = D, 0); (135^0 = D, D)$, where D represents the distance of analysis, in terms of pixels.

T1-weighted 2D features		
GLCM	Entropy	0,2; 2,2; 2,-2; 3,0; 0,3; 3,3; 3,-3; 0,4; 4,0; 4,4; 4,-4;
	Sum Entropy	1,0; 0,1; 1,1; 1,-1; 2,0; 0,2; 2,2; 2,-2; 3,0; 0,3; 0,4;
	Ang. Sec. Moment	0,2; 2,2; 2,-2; 3,0; 0,3; 3,3; 3,-3; 4,0; 4,4; 4,-4;
T2-weighted 2D features		
GLCM	Entropy	1,0;,0,1;,1,1;,1,-1;,2,0;,0,2;,2,2;,2,-2; 3,0;,0,3;,3,3;,3,-3;,4,0;,0,4;,4,4;,4,-4;
	Difference Entropy	0,1; 0,2; 0,3; 3,3; 0,4; 4,4;
	Ang. Sec. Momen	0,2; 2,2; 0,3; 3,3; 3,-3; 4,0;
	Difference Variance	0,1; 1,-1; 0,2; 2,-2; 0,3; 3,-3; 4,0; 0,4;
	Contrast	0,1; 1,-1; 0,2; 2,-2; 0,3; 3,-3; 0,4; 4,4; 4,-4;
	Correlation	0,1;1,-1; 0,2; 2,-2; 0,3; 3,-3; 4,0; 0,4;
	Inverse Diff. Moment	0,1; 1,1; 1,-1; 0,2; 2,2; 2,-2; 3,3; 4,4;
GLRLM	RL Non Uniformity	Horizontal
	Short Run Emphasis	Horizontal; Vertical; 45^0
	Long Run Emphasis	Vertical; 45^0 ; 135^0
	Fraction	Vertical; 45^0; 135^0
Histogram		Max, Variance, 90^{th} Perc, 99^{th} Perc

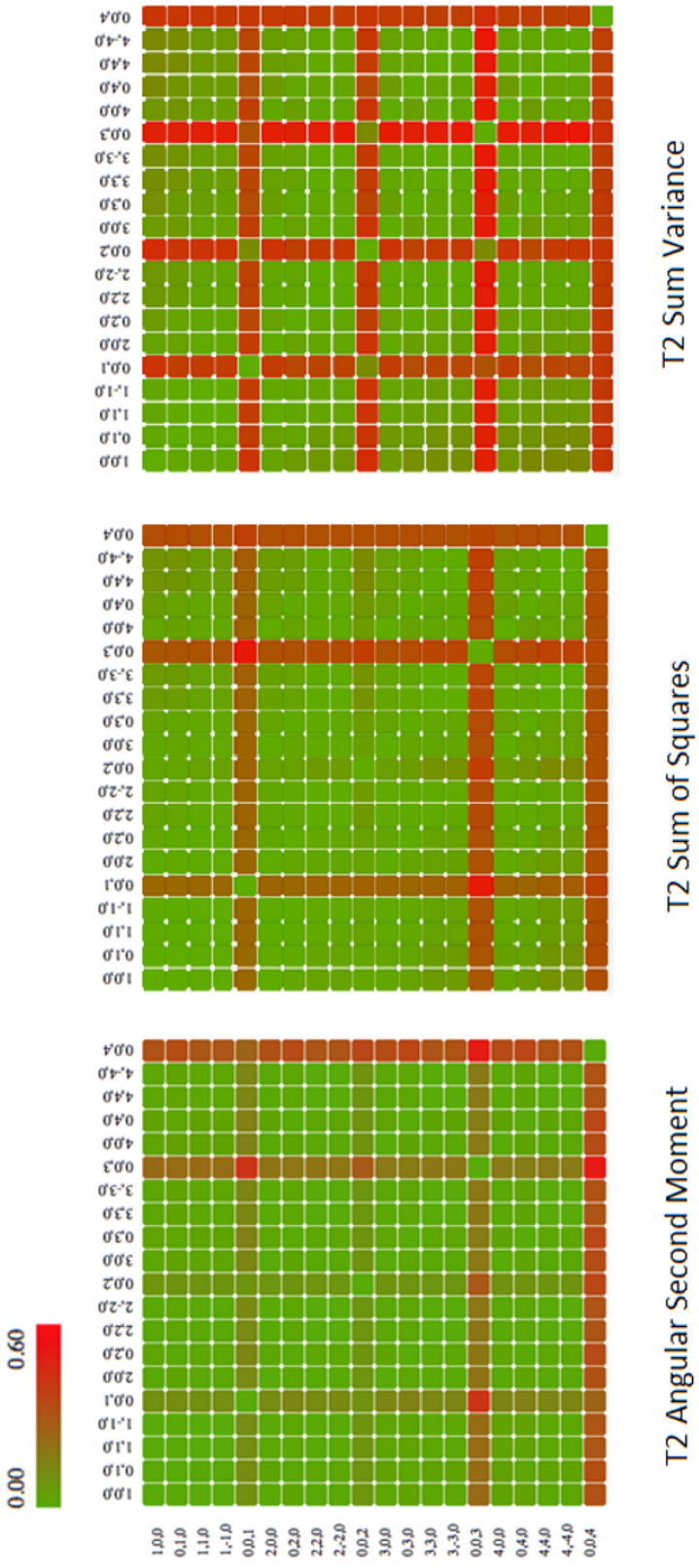


Figure 5.2: Distance maps based on Pearson Correlation metric, measured for variations of three features: T2 Angular Second Moment, Sum of Squares and Sum Variance. The distance was calculated as 1 - r, where r is the Person correlation. A distance of zero would therefore be a correlation of 1.

Table 5.4: A table showing a summary of the T1 and T2-weighted 3D features chosen by entropy-MDL during the feature selection stage. Note that 3D GLCM features are represented using a three-digit offset rather than a two-digit one, where the third digit identifies whether the analysis took place along the z-axis (between slices).

T1-weighted 3D features		
GLCM	Sum of Squares	1,1,0; 1,-1,0; 1,0,0; 0,1,0; 2,0,0; 0,2,0; 0,0,2; 3,0,0; 0,3,0; 0,0,3; 0,4,0; 4,-4,0;
	Sum Average	0,0,2;
	Sum Variance	1,0,0; 0,1,0; 0,0,2;
	Entropy	0,0,3;
	Difference Entropy	0,0,3;
	Volume	1,0,0; 0,1,0; 1,1,0; 1,-1,0; 0,0,1; 2,0,0; 0,2,0; 2,2,0; 2,-2,0; 3,0,0; 0,3,0; 3,3,0; 3,-3,0; 0; 0;3; 4,0,0; 0,4,0; 4,4,0; 4,-4,0;
GLRLM	GL Non-Uniformity	Horizontal; Vertical; 45 degrees
	RL Non-Uniformity	45 degrees; 135 degrees
Histogram		Kurtosis
T2-weighted 3D features		
GLCM	Correlation	0,1,0; 1,1,0; 1,-1,0; 0,0,1; 0,2,0; 2,2,0; 0,3,0; 3,3,0; 3,-3,0; 0,4,0; 4,4,0; 4,-4,0;
	Contrast	0,1,0; 1,1,0; 0,0,1; 0,2,0; 2,2,0; 2,-2,0; 3,0,0; 0,3,0; 3,3,0; 3,-3,0; 4,0,0; 4,4,0; 4,-4,0; 0,4,0;
	Inverse Difference Moment	0,0,1; 2,0,0; 2,-2,0; 0,0,2; 3,0,0; 3,-3,0; 4,-4,0;
	Entropy	1,0,0; 1,1,0, 1-,1,0; 0,0,1; 2,2,0; 2,-2,0; 3,3,0; 3,-3,0;
	Sum Average	1,0,0; 0,0,2;
	Sum Variance	0,1,0; 0,2,0; 0,3,0; 3,3,0; 0,4,0; 4,-4,0;
	Difference Variance	0,1,0; 1,-1,0; 0,2,0; 2,2,0; 2,-2,0; 0,3,0; 3,-3,0; 0,4,0; 4,4,0; 4,-4,0;
	Difference Entropy	0,1,0; 0,0,1; 0,2,0; 2,2,0; 0,3,0; 3,3,0; 3,-3,0; 4,0,0; 0,4,0; 4,4,0; 4,-4,0;
GLRLM	Short Run Emphasis	Vertical; 45 degrees
Histogram		Min, Max, Mean, Variance, Kurtosis, 50th Perc, 90th Perc, 99th Perc, Gradient Kurtosis;

5.3.3 Classification Results and Statistical Findings

AUC values obtained with 2D and 3D features on LOOCV are depicted for each of the classifiers in Figure 5.3. The AUC has an important statistical property: it is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Thus, comparing the obtained AUC values can give an insight to the overall classification performance of classifiers. Table 5.5 gives detailed outcomes of the LOOCV results, in terms of accuracy, sensitivity and specificity for each of the three tumour types.

By inspecting the results illustrated in Figure 5.3 and in Table 5.3, one can see that the use of three-dimensional textural features generally enabled classifiers to capture more information about the tumours and consequently lead to improvements in classification accuracy, sensitivity and specificity. For instance, SVMs overall AUC improved by 13% and it was able to classify MB, PA and EP with a sensitivity increase of 14%, 20% and 28% respectively. SVM specificity performance also increased for MB, PA and EP by 18%, 4% and 7% respectively. It is worth noting that SVM classifier showed the most improvement in overall performance when comparing 3D with 2D results (13% increase in AUC).

Figure 5.4 shows a scatter plot of two 3D features used to train SVM classifier, namely T1 Sum of Squares (0,0,3) and T2 Sum Average (0,0,2). In order to depict how SVM performed on each data point, actual tumour class was represented by colour and predicted class was represented by point shape. By inspecting the figure, one can see that two EPs, one MB and one PA were misclassified by SVM. Figure 5.5 shows another scatter plot of the same data points, using a different combination of 3D features (T2 Sum Variance (0,1,0) and T2 Skewness).

In terms of statistical significance, the results obtained by McNemar's test suggest that there were significant improvements in classification performance when

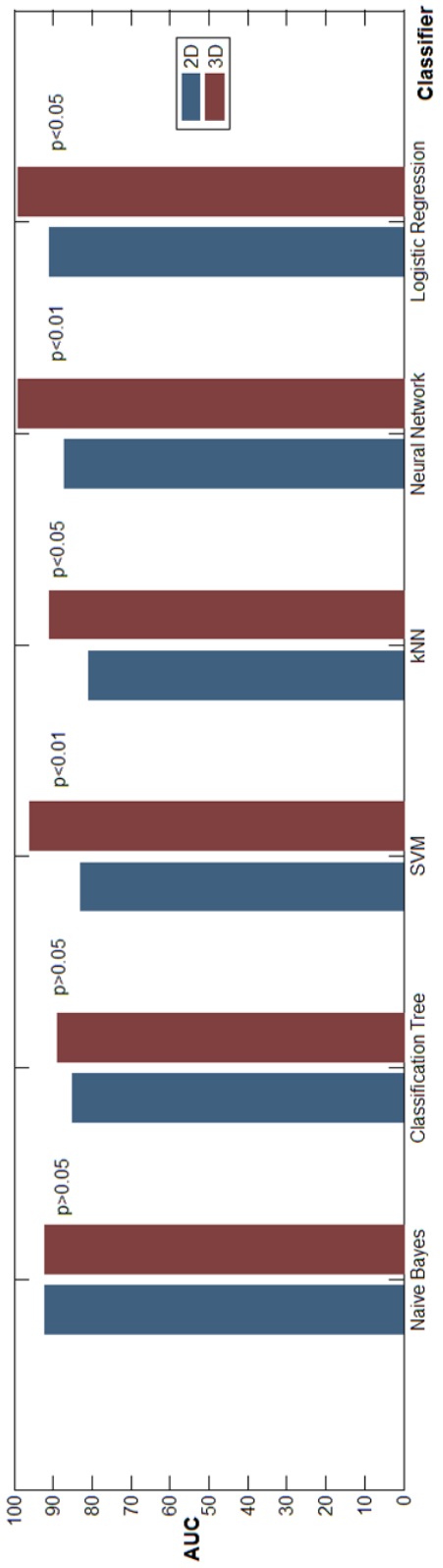


Figure 5.3: Bar charts summarising AUC values obtained with 2D and 3D features on LOOCV for each of the classifiers. The p values obtained with McNemars test comparing 2D and 3D performance are shown for each classifier.

Table 5.5: Summary of classification results obtained by leave-one-out cross-validation (LOOCV) on 2D and 3D textural features. Area Under the ROC Curve, Accuracy, Sensitivity and Specificity are denoted as AUC, Acc, Sens and Spec respectively. Variance of over-all accuracy was calculated, with the assumption of a Binomial approximation to the count of correct classification, as p(1-p)/N, where p is the probability of correct classification and N is the number of samples.

Feature Set	Algorithm	Overall AUC%	MB			PA			EP			Overall Acc%	Var
			Acc%	Sens%	Spec%	Acc%	Sens%	Spec%	Acc%	Sens%	Spec%		
2D	NB	92	90	95	85	88	80	93	90	57	95	83	0.0029
	Tree	85	85	86	85	79	70	86	85	57	90	75	0.0039
	SVM	83	81	86	78	83	70	93	81	43	88	73	0.0041
	kNN	81	79	81	78	81	70	89	81	43	88	71	0.0043
	ANN	87	81	86	78	77	70	82	88	43	95	73	0.0041
	Logreg	91	83	91	78	79	70	86	88	42	95	75	0.0039
3D	NB	92	94	86	100	90	95	86	92	71	95	88	0.0022
	Tree	89	90	96	85	85	74	93	92	71	95	83	0.0029
	SVM	96	98	100	96	94	90	97	92	71	95	92	0.0015
	kNN	91	92	100	85	90	79	97	95	43	93	83	0.0029
	ANN	99	96	100	92	96	90	100	92	71	95	92	0.0015
	Logreg	99	96	100	92	92	85	97	92	71	95	90	0.0019

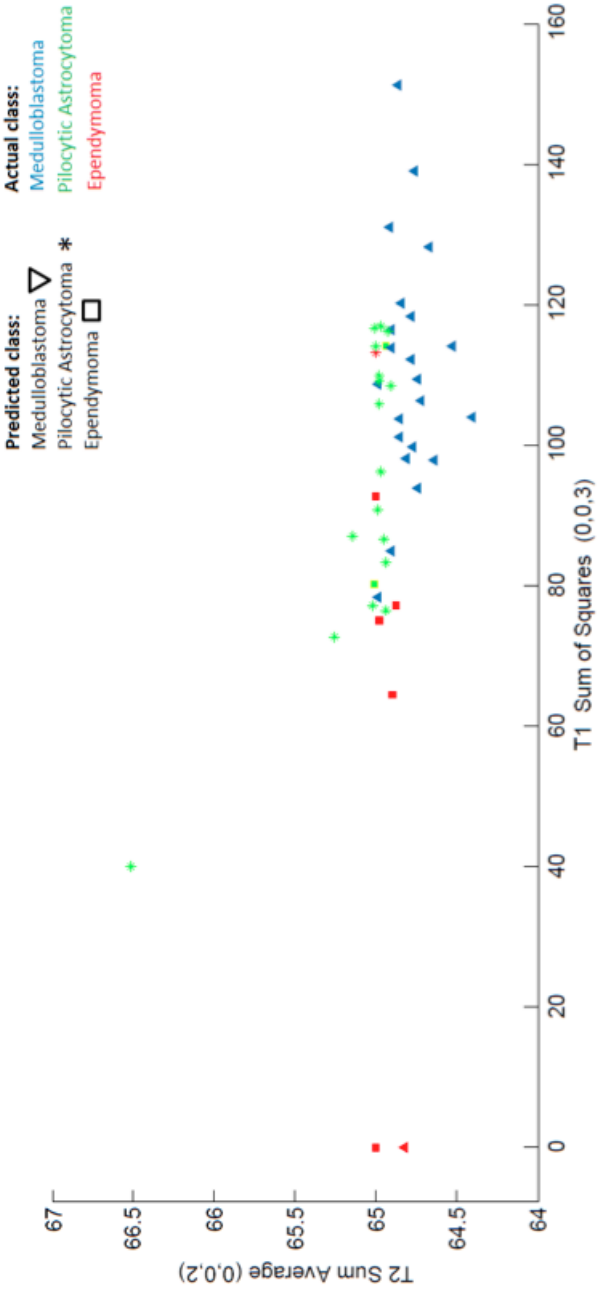


Figure 5.4: Shows a scatter plot of two 3D features used to train SVM classifier, namely T1 Sum of Squares (0,0,3) and T2 Sum Average (0,0,2). In order to depict how SVM performed on each data point, actual tumour class was represented by colour and predicted class was represented by point shape.

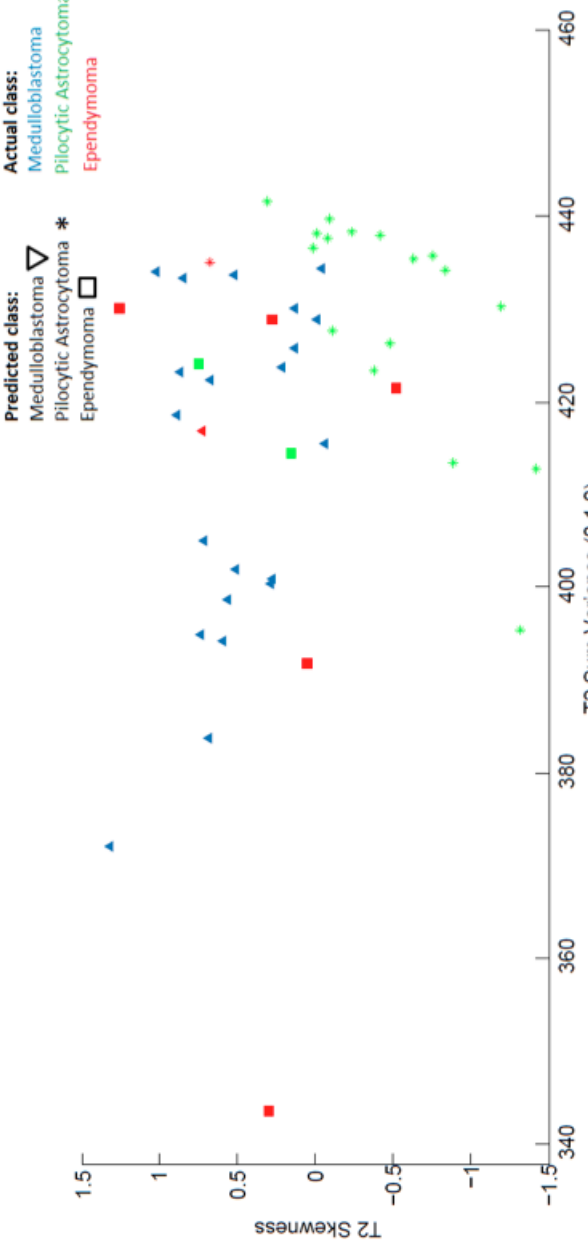


Figure 5.5: A scatter plot of two 3D features used to train SVM classifier, namely T2 Sum Variance (0,1,0) and T2 Skewness. In order to depict how SVM performed on each data point, actual tumour class was represented by colour and predicted class was represented by point shape.

Table 5.6: Contingency table constructed from 3D and 2D-trained SVM classifiers for performing McNemar's test.

		2D-trained SVM		Totals
		Correctly Classified	Incorrectly Classified	
3D trained SVM	Correctly Classified	35	9	44
	Incorrectly Classified	0	4	4
	Totals	35	13	48

Table 5.7: Table summarising results obtained by McNemar's test to assess whether 3D and 2D trained classifiers showed significant differences in performance.

Classifier	p
NB	0.5 (>0.05)
Tree	0.3 (>0.05)
SVM	0.004 (<0.01)
kNN	0.021 (<0.05)
ANN	0.004 (<0.01)
Logreg	0.015 (<0.05)

3D features were used by SVM, kNN, ANN and logistic regression. The obtained two-tailed p values were <0.01, <0.05, <0.01 and <0.05 respectively. Whilst there was an overall improvement demonstrated by Classification Tree (for e.g. 4% improvement in AUC), the results obtained by McNemar's test suggest that this improvement is not statistically significant ($p = 0.3$). Bayesian classifier did not demonstrate improvements in performance when trained with 3D features. A summary of the results obtained with McNemar's test to compare 3D and 2D performances on LOOCV is available in Table 5.7.

With regards to the performance of individual classifiers with 3D features, the highest AUC values were obtained by logistic regression and neural network classifiers (99%), and the lowest were obtained by classification tree (89%). An

interesting finding was that pairwise comparisons carried out by McNemars test did not show logistic regression and neural network to statistically outperform the rest of the classifiers (p >0.05), thus suggesting that the choice of classification algorithm is not of substantial importance.

Table 5.8: Confidence intervals for overall classification accuracies, obtained by a bootstrapping of samples 1000 times. 2D and 3D textural features were used. The lower and upper bounds were chosen to be the accuracies at positions 25 and 975 of the sorted accuracies list, respectively.

Algorithm	2D		3D	
	Lower %	Upper %	Lower %	Upper %
NB	58	92	67	96
Tree	58	92	63	96
SVM	63	96	71	96
kNN	63	96	71	96
ANN	73	96	67	96
Logreg	58	96	71	96

The confidence intervals generated by bootstrapping are listed in Table 5.8. Our overall classification accuracies reported in Table 5.5 fall within the calculated confidence intervals, suggesting that there need not be reasons for concern with regards to potential over-fitting.

T1 and T2-weighted features were analysed independently and tested with neural network classifier using the LOOCV scheme. This was done to get an idea to whether optimal performance can potentially be achieved using a single modality, thus simplifying model complexity and reducing computational costs. Whilst the concatenation of 2D features did not yield improvements in performance, the results obtained with 3D features suggest otherwise. 99% AUC was obtained with concatenated 3D features, as opposed to 90% and 88% on T1 and T2 respectively. A bar chart that summarises the obtained AUC values when T1 and T2-weighted features were tested independently is shown in Figure 5.6. These findings sup-

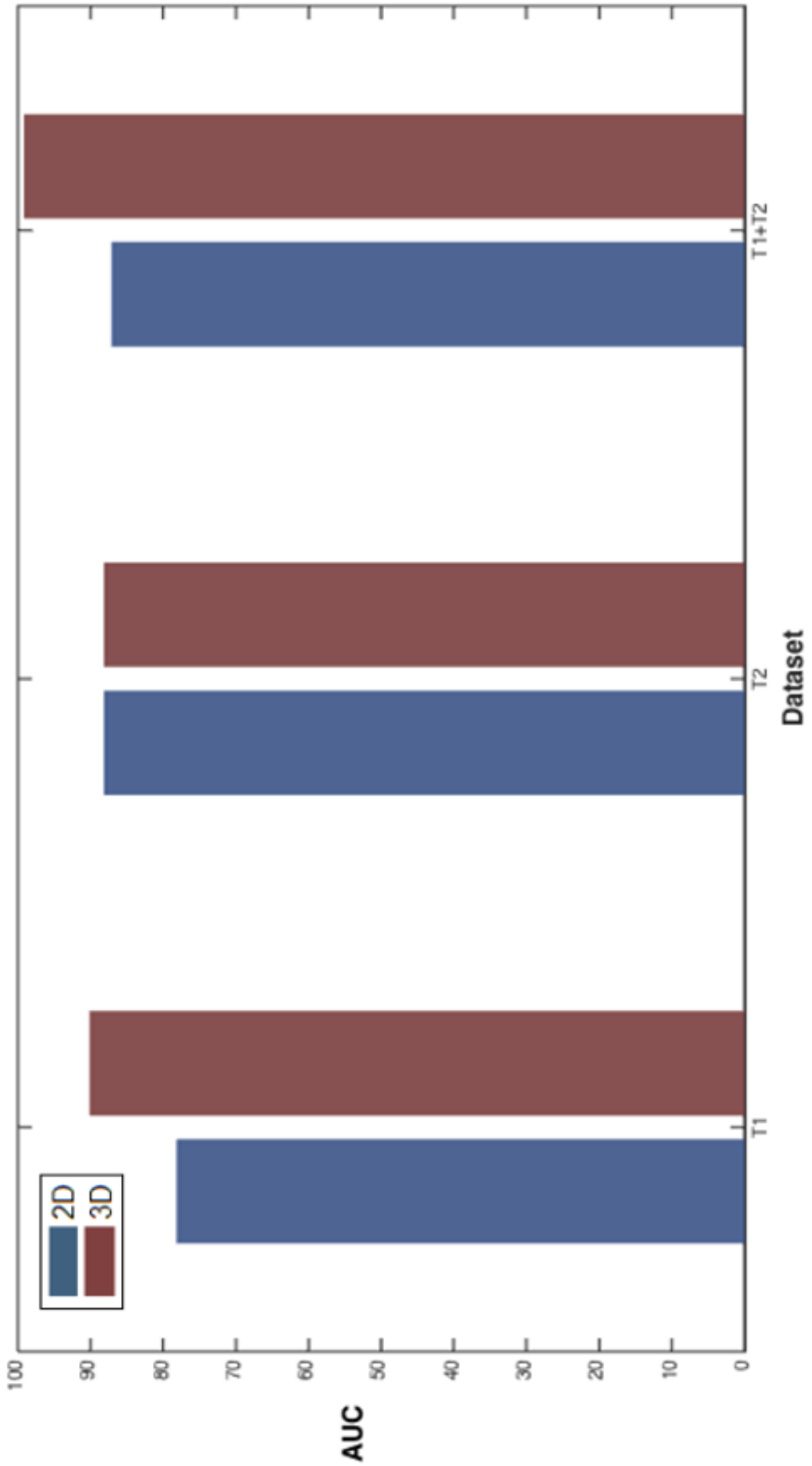


Figure 5.6: A bar chart that summarises the obtained AUC values when T1 and T2-weighted features were tested independently using LOOCV on the 2D and 3D datasets. 99% AUC was obtained with concatenated 3D features, as opposed to 90% and 88% on T1 and T2 respectively. These findings support the use of both T1 and T2-weighted 3D features for optimal classification performance.

port the use of both T1 and T2-weighted 3D features for optimal classification performance.

The classification results obtained by stratified ten-fold cross-validation on the selected 2D and 3D features are summarised in Table 5.9. The results show very similar patterns to those obtained with LOOCV and therefore provide additional reassurance.

Table 5.9: Summary of classification results obtained by stratified 10-fold cross-validation on 2D and 3D textural features. Area Under the ROC Curve, Accuracy, Sensitivity and Specificity are denoted as AUC, Acc, Sens and Spec respectively. Variance of over-all accuracy was calculated, with the assumption of a Binomial approximation to the count of correct classifications, as p(1-p)/N, where p is the probability of correct classification and N is the number of samples.

			MB			PA			EP				
Feature Set	Algorithm	AUC %	Acc %	Sens %	Spec %	Acc %	Sens %	Spec %	Acc %	Sens %	Spec %	Overall Acc %	Var
2D	Naive Bayes	92	88	91	85	88	85	89	92	67	98	83	0.0029
	Classification Tree	86	88	86	89	83	80	86	88	67	93	74	0.004
	SVM	85	81	86	78	83	70	93	85	57	90	73	0.0041
	kNN	88	79	81	78	83	70	93	83	57	88	73	0.0041
	ANN	96	83	91	78	81	70	89	90	57	96	77	0.0037
	Logistic Regression	96	83	91	78	81	70	89	90	57	96	77	0.0037
3D	Naive Bayes	93	94	86	100	90	95	86	92	71	95	88	0.0022
	Classification Tree	88	85	86	85	85	74	93	88	71	90	80	0.0033
	SVM	95	96	96	96	94	95	93	94	71	98	92	0.0015
	kNN	93	88	91	85	88	79	93	88	57	93	82	0.0031
	ANN	99	92	91	92	92	90	93	92	71	95	88	0.0022
	Logistic Regression	99	96	100	92	92	84	97	92	71	95	90	0.0019

5.3.4 PCA-based Results

The findings obtained with PCA-based pipeline show a poorer performance when compared to Entropy-MDL, for Naive Bayes, Classification Tree, SVM and kNN classifiers (AUC ranged between 61% and 84%). However, neural network and logistic regression were able to yield comparable AUC values that ranged between 92% and 95%. PCA-based results are summarised in the form of a bar plot, as per Figure 5.7.

Scatter plots of PC1 vs. PC2 are shown in Figures 5.8 and 5.9. Whilst PCA tends to be used in a black box manner throughout the supervised learning literature [65], [66], it would be interesting to explore whether features deemed as important by PCA match findings by Entropy-MDL. This could perhaps be done by back-projecting PC weights to original feature values. This is, however, not of immediate interest within the scope of this experiment, and thus we focus on assessing whether reducing dimensionality using PCA would enable classifiers to perform more accurately. Nevertheless, detailed comparison of PCA to the supervised approach implemented throughout this thesis would be an interesting future extension of this research.

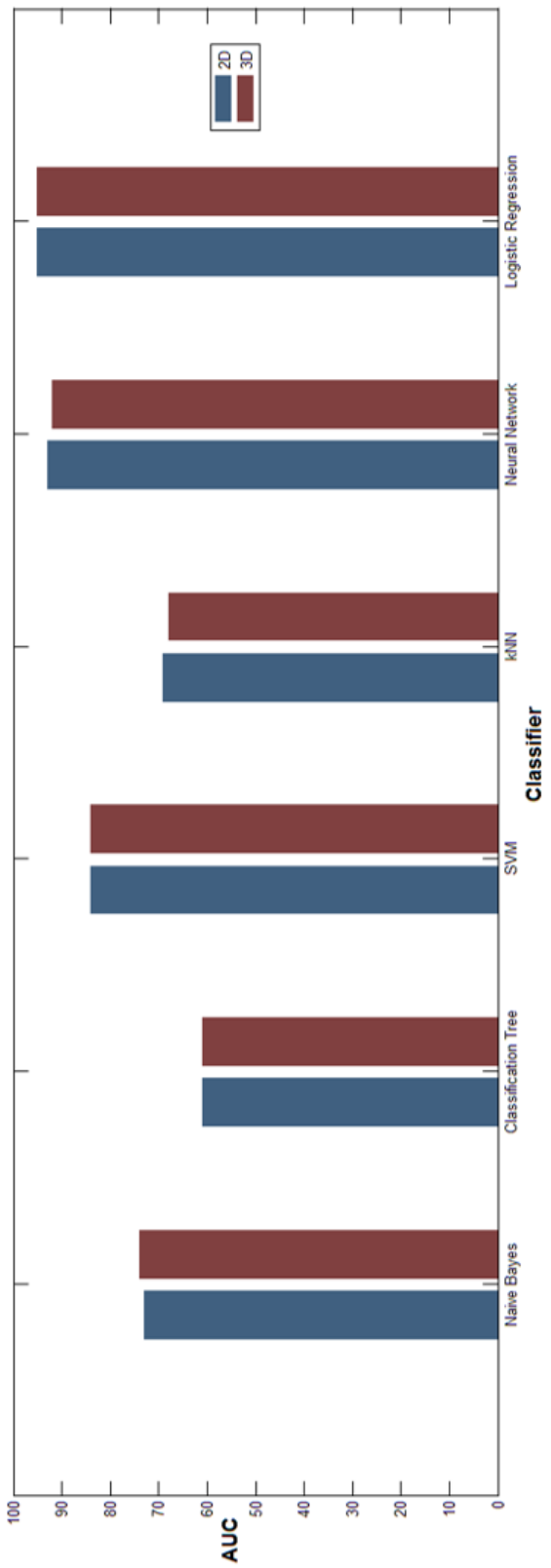


Figure 5.7: A bar plot of AUC results obtained with the PCA-based pipeline. LOOCV scheme was used. One could see how the results obtained with nave bayes, classification tree, SVM and kNN show a poorer performance when compared to entropy-MDL (AUC ranged between 61% and 84%). However, neural network and logistic regression were able to yield comparable AUC values to those obtained with entropy-MDL (92% to 95%).

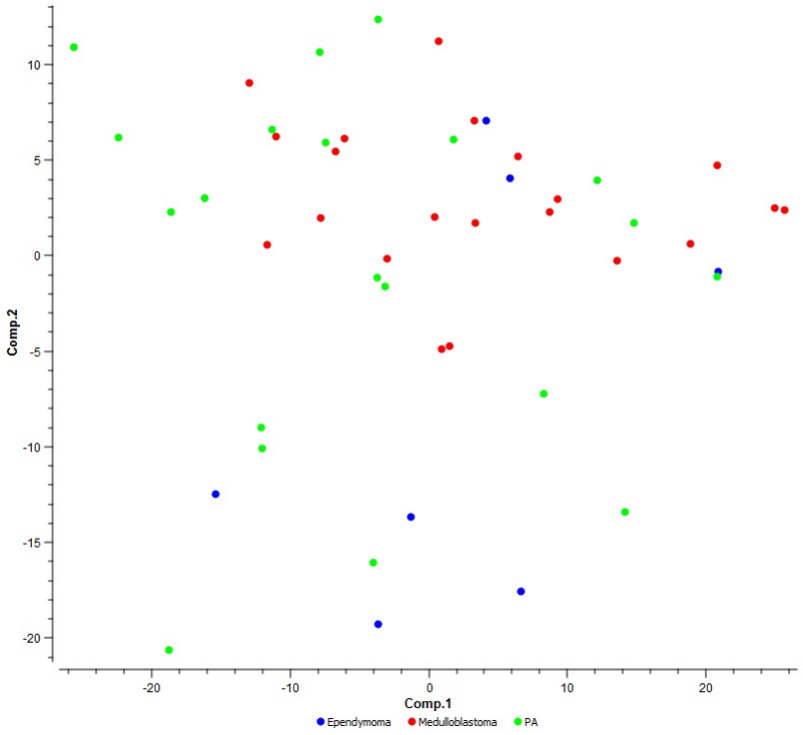


Figure 5.8: A scatter plot of PC1 vs. PC2 using 2D features

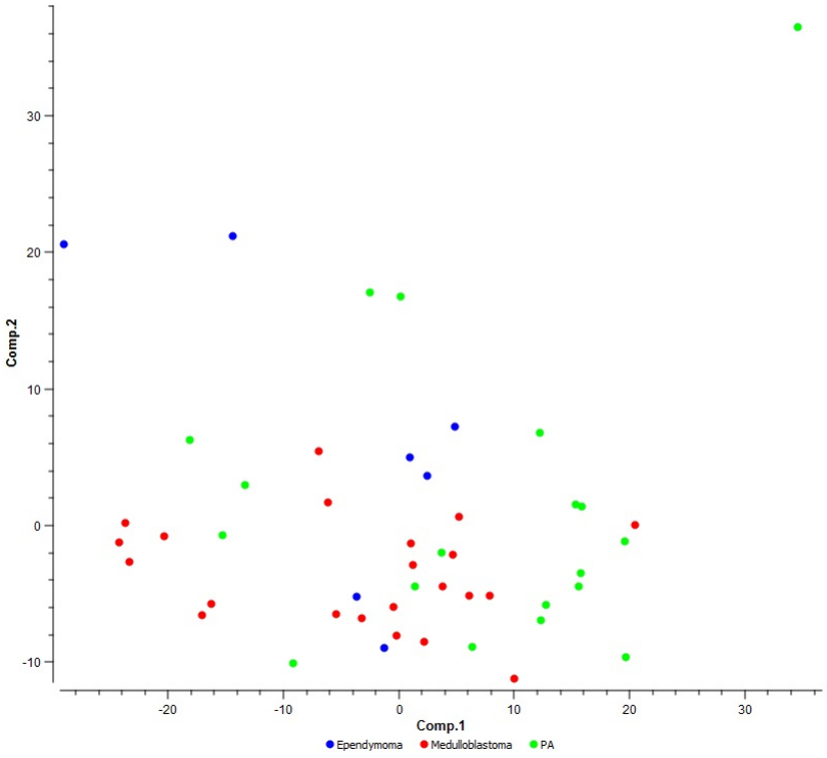


Figure 5.9: A scatter plot of PC1 vs. PC2 using 3D features

5.3.5 Radiological Reporting Benchmark

The total number of reports reviewed was 47, as one report was missing. The diagnoses suggested by the radiologists were recorded and compared to the final diagnoses that were confirmed by histopathological examinations. Reviewing the cases included in the study showed that 22 of the 47 reports (47%) had the correct diagnosis stated. Out of those 22 reports, 8 had an alternative diagnosis proposed. In cases where an alternative tumour type was proposed, one report specifically mentioned their order of likelihood. 17 out of the 47 reports (36%) did not have any differential diagnosis proposed, while 6 out of the 47 reports (13%) had a single incorrect diagnosis. Excluding cases where no diagnosis was proposed gives 20/30 (66%) where the only diagnosis given or the first diagnosis in a list was correct. Overall, some level of uncertainty exists in 27/47 (57%) of the reports. It is worth noting that, as discussed in Chapter 1, a radiologist's job does not require offering a final tumour diagnosis but rather to offer initial characterisation of its appearance; a likely reason for such apparent uncertainty.

5.4 Discussion

Classification results obtained with the conventional 2D TA approach are in line with what the current state-of-the-art has achieved in the paediatric literature. For the commonly reported classification task of separating MB, PA and EP, the accuracies of 83% (SVM) and 87%(ANN) are comparable with the work reported in [36] (71% with T1 and 74% with T2) and [37] (63% EP, 81% PA and 94% MB)

The primary aim of this study was to determine whether the inclusion of multi-slice information obtained through 3D TA of conventional MRI could improve diagnostic classification of childhood brain tumours. The obtained results suggest that the value of TA can be maximised using 3D features in paediatric settings. Statistical findings obtained with McNemar's test indicate that this improvement in performance was significant for four of the six classifiers that were tested. It is worth noting, however, that all six classifiers showed relatively low EP sensitivity. This is likely to be due to the highly imbalanced nature of the dataset, where only 7 EP samples were present in the cohort, leading to the three tumour classes not being equally represented. In other words, the limited number of EP samples seems to have caused the classifiers to categorise most cases as MB and PA. Another possibility is that EP, being typically heterogeneous masses, might have textural properties that are common with the other two tumour types, which could lead to classifiers not being able to accurately discriminate it from the two other classes.

In terms of important features, most of the ones chosen during the feature selection stage are attributes that were derived from GLCM and histogram statistics techniques. An important point to keep in mind is that the feature selection technique used is supervised, in the sense that it requires prior knowledge of the class label. This means that an element of over-optimistic bias might have been introduced during the classification model validation stage, as feature selection

was carried out outside the leave-one-out loop. However, since the same methodology was used when analysing classifiers trained with both 2D and 3D, any bias would have occurred to both sets of classifiers.

An interesting observation is the directional sensitivity of feature selection towards GLCM and GLRLM features measured across different combinations of pixel directions and distances. Entropy-MDL is a technique used to discretise features by finding a splitting value that yields the best gain in entropy. This is repeated recursively with a stopping criteria that is based on the Minimal Description Length principle. The use of this technique as a feature selection method is based on the assumption that since a feature's entropy can be used as a measure of its discriminative power, those features that were rejected by the algorithm can be assumed to be redundant. A feature would not be discretised if no appropriate cut-off points are found. Therefore, whilst features measured across different directions compute strongly correlated patterns (as shown in Figure 5.2) some of them do not have sufficient information in terms of gain in entropy. The inclusion of these features will therefore not yield extra value to the classification performance.

Another interesting finding is that only twelve features in the highly ranked 3D subset were a result of analysis along the z-axis (inter-slice information). Hence, the addition of inter-slice patterns has contributed to improvements in classification performance, but it is likely that improvements were mostly due to classifiers being able to capture information from the whole volumetric ROI, compared to just selecting a single slice and extracting features that are not representative enough to classify tumours. The limited number of important through-plane features maybe due to the presence of slice gaps, which ranged between 0.8-1.5 mm for T1 and 0.6-1.5 mm for T2 in the dataset we used.

With regards to the results obtained with PCA, the reduced overall perfor-

mance shown by four of the tested classifiers is likely due to the fact that PCA computes new meta-features, namely principal components. These are linear combinations of the original attributes, which may be difficult for classifiers to generalise well with, as the feature space significantly changes and the original features may lose their original meaning. The fact that PCA is a dimensionality reduction and not a feature selection tool adds the additional limitation of not being able to deduce a definite sub-set of important features, making it an impractical option for understanding relationships between the original textural features and classification performance.

The exact accuracy of radiological reporting is unknown and the study presented here suffers from the disadvantage that the radiologists did not have to offer a diagnosis or even a differential diagnosis[2]. Despite this limitation, the review has the advantage that it was contemporaneous and gives some insight into the difficulties of radiological reporting. If only the reports where a single correct diagnosis is offered are taken as correctly diagnosed, the overall success rate is 14/47 (30%). However, if we exclude cases where no diagnosis was proposed, the accuracy is 14/30 (47%) and if the first diagnosis in a list is taken as the favoured one, 20/30 (66%) are correct. Whilst the accuracy of diagnosis for cases where no diagnosis was proposed could be greater than this, it would seem unlikely. In reality, the most common reason for not offering a diagnosis is uncertainty and it is an interesting observation that some level of uncertainty exists in 27/47 (57%) of the reports. It may well be that a key role of a decision support system based on texture analysis is to improve this uncertainty. To illustrate how the use of TA can help achieve this, consider Figure 5.10, which shows a summary of probabili-

[2]Conventionally, radiologists produce an initial characterisation of the tumour's appearance on the basis of a combination of their training, experience and individual judgement. The radiologist's job is not to offer a final diagnosis as the current gold-standard is histopathological examination.

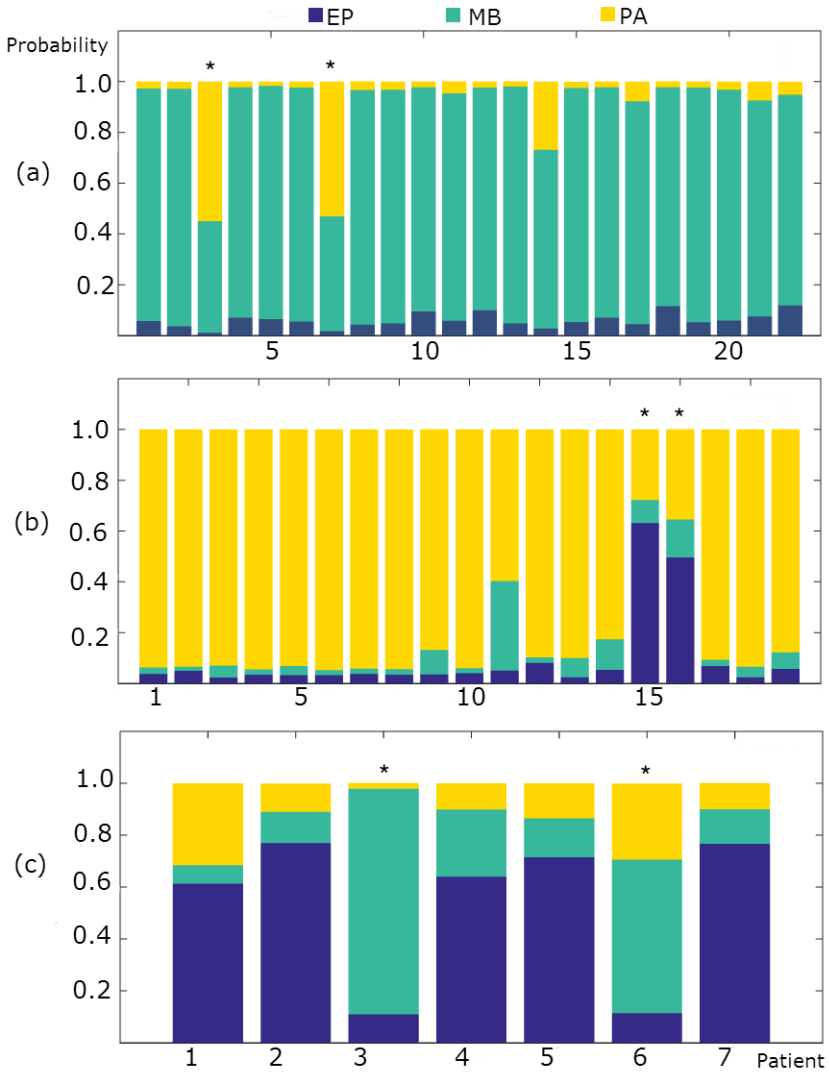


Figure 5.10: A bar plot summarising probabilities assigned to each individual diagnosis by the neural network classifier during LOOCV. Actual class is (a) Medulloblastoma (b) Pilocytic Astrocytoma and (c) Ependymoma. Note that bars marked with an asterisk indicate a misclassification made by the classifier.

ties assigned to each individual diagnosis by the neural network classifier. Taking Figure 5.10(a) as an example, one can see how the two misclassified MB samples (marked with an asterisk) had close likelihoods of being MB and PA, according to the classifier; suggesting limited confidence in the final diagnosis. Such information could be potentially valuable as diagnostic aids for radiologists in practical settings.

5.5 Study Limitations and Future Work

In terms of study limitations, the use of a small cohort of 48 patients was the main one. Whilst two cross-validation schemes were used, in addition to carrying out bootstrapping in order to provide reassurance of results consistency, the use of a large cohort would help confirm the robustness of TA. A multicentre study that includes datasets obtained from different scanners is therefore the next step of this research. A multicentre study would also be a robust way of assessing cross-centre transferability of textural features. For instance, feature selection could be carried out on data obtained from one centre, followed by testing those features that were shown to be important on data from another centre.

Within the scope of this thesis, this limitation was adressed through a multicentre study, which is discussed in the next chapter.

5.6 Conclusion

In conclusion, the study presented in this chapter demonstrated how texture analysis of pre-contrast T1 and T2-weighted MR images and machine learning algorithms could be used to design quantitative models for objective evaluation of common paediatric brain tumours. An essential outcome of this study is that 3D

features, combined with supervised classification methods, achieved improved classification performance compared to 2D features. Regarding feature importance, those attributes derived from GLCM and histogram statistics have the highest discriminating potential. The experiment presented here provides a foundation for the use of 3D TA methods to build intelligent computational tools that can help achieve early diagnosis in paediatric oncology. Benefits of such tools would include reducing surgical procedures, improving surgery and therapy planning and supporting discussions with the patients' families. Future work will focus on examining the robustness of 3D TA by extending this study to multicentre cohorts. Additionally, it will be interesting to look into appropriate class balancing techniques in order to mitigate the lack of enough EP samples in the population used. The work presented here meets the first objective of this thesis: to carry out a practical investigation on the diagnostic efficacy of MRI TA.

Chapter 6

A Multicentre Investigation on the Transferability of TA

Some aspects of the work presented here were published in [**P04**] and [**P05**]. Publication details can be found on Page *xx*.

6.1 Introduction

Despite the positive results reported in the adult and childhood brain MRI literature, TA has not yet found its way into routine clinical practice. This is perhaps due to the sensitivity of textural features to variations in MR acquisition parameters, which may impede the transfer of results across various imaging centres [71]. In addition to this, the efficacy of TA is heavily dependent on the choice of textural features used to capture imaging patterns, which is linked to the choice of feature-selection methods used [71]. However, very little comparative work is available on studying the aforementioned issues [71].

The study presented here expands the work discussed in Chapter 5 to include multicentric datasets obtained from three different hospitals across the UK. The primary aim of this study was to determine the efficacy and cross-centre transferability of 3D TA for non-invasive classification of childhood brain tumours from MR images. The study also aimed to investigate, through the use of supervised feature selection, the nature of features that are most likely to train classifiers that can generalise well with the 3D textural data. Finally, the issue of class imbalance, which arises due to some tumour types being more common than others, was looked into. To the best of the author's knowledge at the time of writing, there are no published studies that used multicentre cohorts in order to assess the effectiveness and transferability of 3D MRI TA in paediatric oncology.

6.2 Materials and Methods

6.2.1 Cohort Details and Image Acquisition

The clinical material used in this retrospective study consisted of pre-contrast T1 and T2-weighted MR images of 134 children with verified and untreated brain tumours. 45 were MB, 71 were PA and 18 were EP. Image acquisition was carried out at three centres: Birmingham Children's Hospital (BCH), Nottingham University Hospital (NUH) and Great Ormond Street Hospital (GOSH).

The following scanners were used for image acquisition: 1.5 T Siemens Symphony, 1.5 T Siemens Avanto (Siemens Healthcare, Erlangen, Germany), 1.5 T General Electric Signa (GE Healthcare, Little Chalfont, UK), 1.5T Phillips Intera and 3 T Phillips Achieva (Philips Healthcare, Amsterdam, Netherlands), following a common protocol defined by the Children's Cancer and Leukaemia Group (CCLG) Functional Imaging Group. All images were anonymised and held at a secure e-repository [4] provided by CCLG, from which the data was downloaded for use in this study.

Table 6.1: A table summarising models and field strengths for the three centres.

BCH	NUH	GOSH
GE Signa 1.5 Tesla	Phillips Achieva 3.0 Tesla	Siemens Avanto 1.5 Tesla
Siemens Symphony 1.5 Tesla	Phillips Intera 1.5 Tesla	Siemens Symphony 1.5 Tesla

6.2.2 Image Pre-processing

In keeping with the methodology used in the single-centre study presented chapter 6, image pre-processing was carried out by manually selecting axial slices from the dataset, segmenting the tumour using the Snake GVF algorithm, and normalising the images using the $\mu + / - 3\sigma$ technique.

6.2.3 Extraction of Textural Features

MaZda software was used to carry out 3D TA based on the histogram statistics, absolute gradient, GLCM and GLRLM techniques, extracting the same features that were used in the single centre study.

6.2.4 Feature Selection

The efficacy of TA is heavily dependent on the choice of textural features used to capture imaging patterns, which is linked to the choice of feature-selection methods used. Hence, a number of feature selection algorithms were considered in study, the first being ReliefF [56]. Entropy minimum descriptive length (MDL) discretisation technique, which was used in the single-centre study reported in Chapter 5, was also considered here [57]. Thirdly, we were also interested in studying the use of a feature selection pipeline, comprising a hybrid of both algorithms: Entropy-MDL and ReliefF, to see whether their combined use could provide additional classification value.

6.2.5 Classification Model

Using python's Orange library, a cost-based support vector machine (C-SVM) classifier was used, using RBF kernel function and a cost coefficient (C) of 1, to be trained with textural features.

6.2.6 Model Validation

Examining Transferability by Pairwise Testing on Unseen Data

To determine the practical influence of differences in textural feature-sets extracted from different MRI centres, three different instances of the SVM classifier were

created, each being trained on features extracted from one of the three hospitals. Testing was carried out by assessing how each SVM performed on unseen datasets, which were obtained from the other two hospitals. For example, the performance of an SVM trained with Birmingham Children's Hospital data was evaluated by testing on datasets obtained from Great Ormond Street Hospital and Nottingham University Hospital.

Evaluation of classification performance was carried out by measuring the area under the ROC curve, or simply AUC. The training and testing process was performed separately for Entropy-MDL, ReliefF and the hybrid pipeline. Note that the reported results were obtained after the feature selection algorithms were optimised, by steadily increasing the percentage of chosen ranked features until optimal classification performance was yielded.

Estimating Overall Performance Using LOOCV

Testing the models on unseen data mainly aimed to examine the corss-centre transferrability of TA. In order to get an overall estimate of the models' classification performance, LOOCV was additionally carried out on an aggregated feature-set comprising data from all three hospitals (all 134 samples).

Classification accuracy, sensitivity, specificity and AUC were measured from the results. 95% confidence intervals of classification accuracies were calculated using bootstrapping (1000 samples were generated). Since the aim of this step was not to investigate optimal settings for classification, but to get an estimate of the overall performance, only one feature selection method (Entropy-MDL) was used.

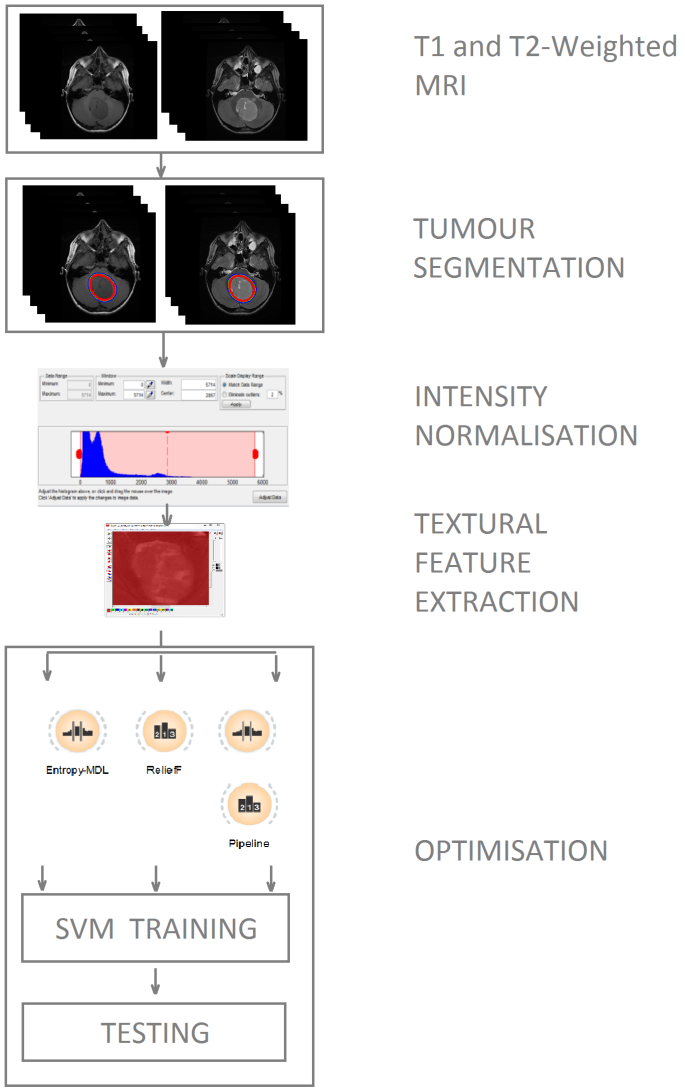


Figure 6.1: A flowchart showing methodological overview of the multicentre experimental set up.

6.2.7 Addressing the Class Imbalance Problem

A dataset is considered imbalanced if the classes are not approximately equally represented. Although the data used for this study had been acquired at three different hospitals, the classes represented are quite imbalanced in the sense that EP forms only 13% of the overall dataset (18/134). This may be problematic because the minority samples might be ignored by the classifier, which could potentially lead to poor EP sensitivity. In order to investigate this, a separate analysis was carried out where the *synthetic minority over-sampling technique* (SMOTE) was applied to the extracted 3D features.

SMOTE was used to create 27 synthetic EP samples by operating in feature space. This method works by taking each minority class sample and introducing synthetic examples along the line segments joining any/ all of the k minority class nearest neighbours. The neighbouring points are randomly chosen depending on the amount of over-sampling required. Using an SVM classifier, LOOCV was carried out on the new feature-set comprising 184 samples. Classification accuracy, sensitivity, specificity and AUC were measured from the test results. 95% confidence intervals of classification accuracies were calculated using a bootstrapping of subjects in the sampling (1000 samples were generated).

6.3 Results

6.3.1 Results from Pairwise Testing on Unseen Data

Table 6.2 and Figure 6.2 show a summary of AUC values obtained with SVM classifier. The mean AUC values obtianed by Entropy-MDL, ReliefF and the hybrid pipeline are 74.5%, 71.8% and 76% respectively. The highest AUC value was obtained when SVM was trained on NUH data and tested on BCH data (86% on ReliefF) - an interesting finding since one of the scanners in NUH uses magnetic field strength of 3T, whereas both scanners used to acqiure BCH data are 1.5T.

For each feature selection method, the number of chosen features that were required to yield optimum AUC are reported in Table 6.3. Whilst with entropy-MDL no manual definition of feature percentages is performed for the algorithm to operate, we reported the number of features that were discretised and hence deemed important by the algorithm.

The optimal features identified for each of the six tests are reported in Tables 6.4 and 6.5. It is worth noting that there was a common set of attributes that were deemed as optimal in all six tests (e.g. *sum of squares, sum average* and *difference entropy*). However, the particular feature variation (i.e. the specific pixel distance and direction) that was identified as important varied greatly across the six tests. For instance, even though the *sum of squares* attribute was identified as an important feature in both tests 1 and 2, test 1 showed that offsets (0,0,3) and (0,0,3) were important, whereas test 2 showed that offsets (1,0,0),(0,1,0) and (0,0,2) were important.

Table 6.2: Optimal AUC values obtained through pairwise testing for multicentre classification.

Test	Training	Testing	Entropy-MDL	ReliefF	Pipeline	Optimal method
ID						
1	BCH	NUH	74%	83%	73%	ReliefF
2	BCH	GOSH	74%	62%	80%	Pipeline
3	NUH	BCH	74%	86%	74%	ReliefF
4	NUH	GOSH	71%	85%	75%	ReliefF
5	GOSH	BCH	76%	60%	76%	Entropy-MDL / Pipeline
6	GOSH	NUH	78%	55%	78%	Entropy-MDL / Pipeline
Mean AUC value:						
Entropy-MDL: 74.5%						
ReliefF: 71.8%						
Pipeline: 76%						

Table 6.3: A table listing the number of features that were needed to yield optimal AUC.

Test	Training	Testing	ReliefF	Entropy-MDL	Pipeline
ID					
1	BCH	NUH	44	123	43
2	BCH	GOSH	50	123	19
3	NUH	BCH	33	85	28
4	NUH	GOSH	34	85	30
5	GOSH	BCH	100	14	14
6	GOSH	NUH	100	14	14

Table 6.4: A table listing the optimal textural features identified in Tests 1 to 3.

Feature name	**Offset**	
Test 1	**T1**	**T2**
Angular Second Moment	(0,1,0) (0,2,0)(0,0,2) (0,3,0)	(0,4,0)
Contrast	(0,0,4)	
Difference Entropy	(0,0,3) (0,0,4)	(0,2,0)(2,-2,0) (3,-3,0) (0,3,0) (0,4,0)(4,-4,0)
Difference Variance	(0,0,4)	(0,1,0)(0,4,0)
Entropy	(0,0,3) (0,0,4)	(0,0,1)
Fraction		45, 135 degrees, Vertical
Histogram	Skewness	Skewness
Inverse Difference Moment	(0,1,0) (0,2,0) (0,3,0) (0,4,0)	(0,2,0) (0,3,0) (0,4,0)(4,-4,0)
Long Run Emphasis	Vertical	
Short Run Emphasis		135 degrees, Vertical, Horizontal
Sum Average	(0,0,3) (0,0,4)	(0,1,0)
Sum Entropy	(0,0,3) (0,0,4)	(0,0,1)
Sum Of Squares	(0,0,3) (0,0,4)	
Test 2	**T1**	**T2**
Angular Second Moment	(0,0,3)	
Difference Entropy	(0,3,0)	
Entropy	(0,0,3)	
Gradient		NonZeros
Histogram		Max, Min, 50%, 90%, 99%, Mean, Variance, Kurtosis
Inverse Difference Moment	(3,-3,0)	(4,-4,0)
Sum Average		(1,0,0),(0,0,2)
Sum Of Squares	(1,0,0), (0,1,0),(0,0,2)	
Test 3	**T1**	**T2**
Angular Second Moment	(0,0,4)	(0,0,4)
Contrast	(0,0,4)	(0,0,4)
Correlation	(0,0,4)	(0,0,4)
Sum Of Squares	(0,0,4)	(1,0,0) (1,1,0) (0,1,0) (1,-1,0) (0,2,0) (0,0,4)
Inverse Difference Moment	(0,0,4)	(0,0,4)
Sum Average	(0,0,4)	(0,3,0) (0,1,0) (0,2,0) (0,0,4)
Sum Variance	(0,0,4)	(0,0,4)
Sum Entropy	(0,0,4)	(0,0,4)
Entropy	(0,0,4)	(0,0,4)
Difference Variance	(0,0,4)	(0,0,4)
Difference Entropy	(0,0,4)	(0,0,4)
Volume	(0,0,4)	(0,0,4)
Histogram		Skewneess

Table 6.5: A table listing the optimal textural features identified in Tests 4 to 6.

Feature name	Offset	
Test 4	**T1**	**T2**
Angular Second Moment	(0,0,4)	(0,0,4)
Contrast	(0,0,4)	(0,0,4)
Correlation	(0,0,4)	(0,0,4)
Sum Of Squares	(0,0,4)	(0,1,0) (1,0,0) (1,1,0) (1,-1,0) (0,2,0) (0,0,4)
Inverse Difference Moment	(0,0,4)	(0,0,4)
Sum Average	(0,0,4)	(0,1,0) (0,2,0) (0,3,0) (0,4,0) (0,0,4)
Sum Variance	(0,0,4)	(0,0,4)
Sum Entropy	(0,0,4)	(0,0,4)
Entropy	(0,0,4)	(0,0,4)
Difference Variance		(0,0,4)
Difference Entropy	(0,0,4)	(0,0,4)
Volume	(0,0,4)	(0,0,4)
Histogram		Skewness
Test 5	**T1**	**T2**
Correlation		(0,0,1)
Sum Of Squares		(0,0,2)
Inverse Difference Moment		(2,2,0)(2,-2,0)(4,-4,0)
Sum Average		(0,0,2)(0,1,0)(0,2,0)
Sum Variance		(0,1,0)
Difference Entropy	(0,0,1)	(3,-3,0)(4,-4,0)
Histogram		Kurtosis
Test 6	**T1**	**T2**
Correlation		(0,0,1)
Difference Entropy	(0,0,1)	(4,-4,0)
Difference Entropy		(3,-3,0)
Histogram		Kurtosis
Inverse Difference Moment		(2,2,0)(2,-2,0)(4,-4,0)
Sum Average		(0,1,0)(0,2,0)(0,0,2)
Sum of Squares		(0,0,2)
Sum Variance		(0,1,0)

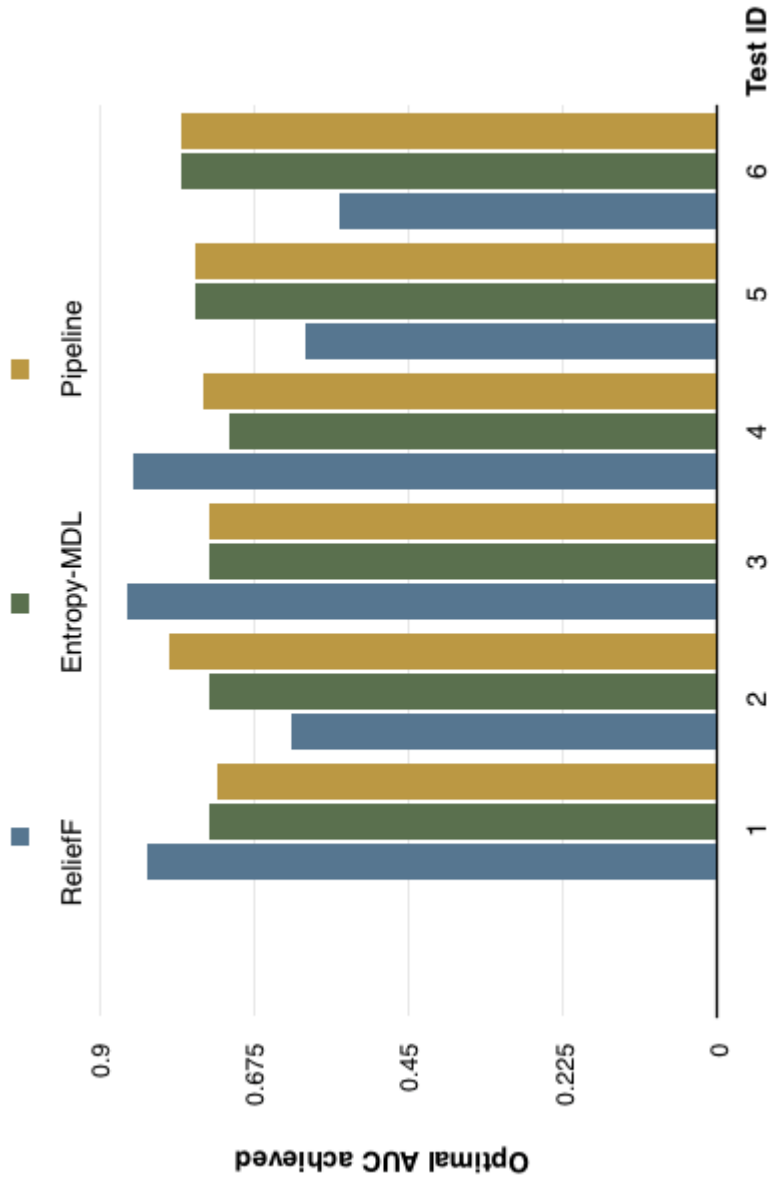


Figure 6.2: Bar chart showing optimal AUC values obtained through pairwise testing for multicentre classification.

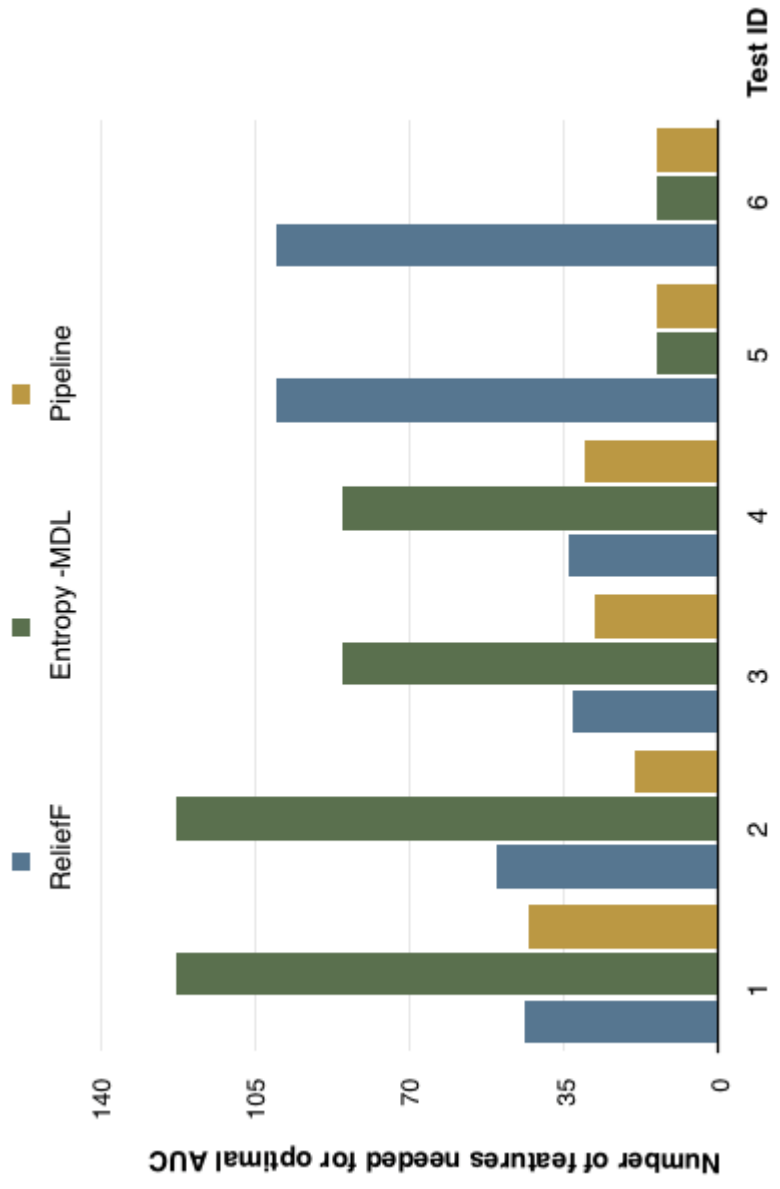


Figure 6.3: Bar chart showing number of features that were needed for optimal AUC.

6.3.2 LOOCV Results

Table 6.6 lists the results obtained when the entire feature-set, comprising all 134 samples, was tested with an SVM classifier using LOOCV. Results were generally satisfying, with the overall AUC being 86%. However, it is worth noting that similar to the results obtained with the single-centre study, EP demonstrated a very low sensitivity value of 11% .

Table 6.6: A table listing the results obtained when the feature-set, comprising data from all three hospitals (134 samples), was tested with an SVM classifier on LOOCV. Entropy-MDL was used for feature selection. Accuracy, sensitivity and specificty are referred to as Acc, Sens and Spec respectively. 95% confidence intervals for the overall classification accuracies were obtained by bootstrapping.

	MB		PA		EP				
AUC	Sens	Spec	Sens	Spec	Sens	Spec	Acc	Variance	95% CI
86%	67%	82%	90%	71%	11%	97%	72%	0.0015	50%-84%

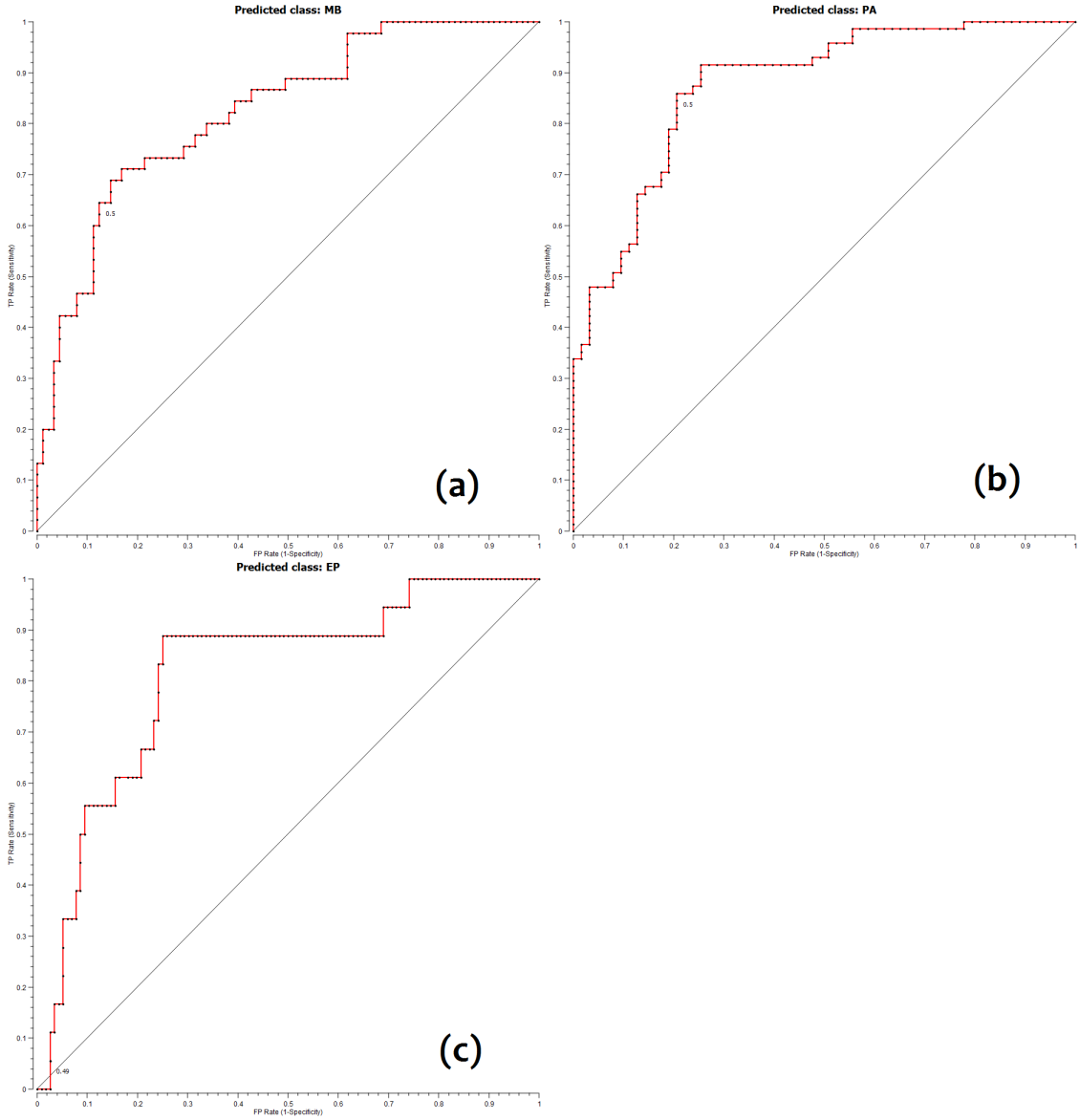


Figure 6.4: ROC curves depicting SVM classifier performance using the LOOCV scheme. All 134 samples obtained from three hospitals were used for the analysis. (a) Medulloblastoma (b) Pilocytic Astrocytoma and (c) Ependymoma. Overall AUC value is 86%.

6.3.3 LOOCV Results After Minority-Oversampling

Table 6.7 lists the LOOCV results obtained when an SVM classifier was used on the new feature-set that comprises an additional 27 (synthetic) EP samples. The noticebale increase in EP sensitivity (from 11% to 87%) suggests that the availability of equally represented classes has enabled SVM to better characterise the data points.

Table 6.7: A table listing the classification results obtained with LOOCV, after SMOTE was applied to generate 27 synthetic EP samples. Entropy-MDL was used for feature selection. Accuracy, sensitivity and specificty are referred to as Acc, Sens and Spec respectively. 95% confidence intervals for the overall classification accuracies were obtained by bootstrapping.

	MB		PA		EP				
AUC	Sens	Spec	Sens	Spec	Sens	Spec	Acc	Variance	95% CI
92%	57%	91%	83%	83%	87%	91%	77%	0.0011	60% - 90%

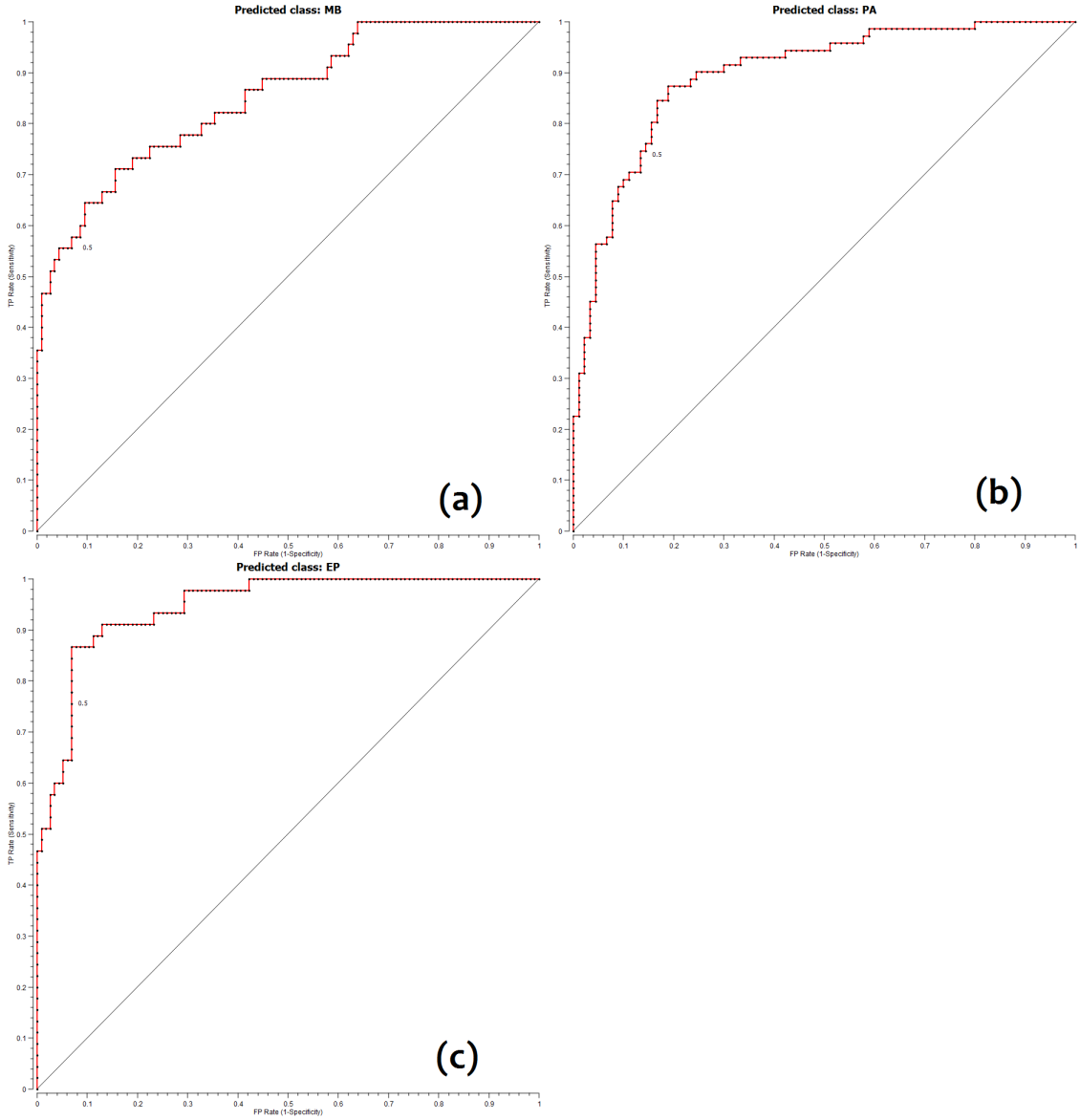


Figure 6.5: ROC curves depicting SVM classifier performance using the LOOCV scheme, after SMOTE was used to generate 27 synthetic ependymoma samples. (a) Medulloblastoma (b) Pilocytic Astrocytoma and (c) Ependymoma. Overall AUC value is 92%.

6.4 Discussion

This chapter presented a multicentre investigation on the efficacy and transferability of using volumetric (3D) statistical textural features extracted from conventional MR images, within a machine-learning framework, to discriminate between the most frequently occurring paediatric brain tumours: medulloblastoma, pilocytic astrocytoma and ependymoma. The study made use of standard pre-contrast T1 and T2-weighted images, which are routinely acquired when children present with suspected brain tumours. For the purpose of this discussion, the two main areas of interest are:

1. Whether the classification results showed enough evidence that 3D TA is a transferrable technique, allowing for its use across multiple centres.

2. The nature of features deduced to be optimal as per feature selection, and how different feature selection methods performed on different tests.

With regards to feature selection, this study looked into comparing the performance of ReliefF, Entropy-MDL and a pipeline comprising both methods. For the three feature-selection methods used, the mean AUC values ranged between 71.8% and 76%. Note that the statistical meaning of AUC can be defined as the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. For each of the tests reported in Table 6.2, optimal AUC ranged between 76% and 86%, which suggests that the use of three-dimensional textural features generally enabled SVM to capture transferable tumour information that could be used to successfully classify images obtained from other imaging centres.

Feature selection results suggest that similar aspects of tumour texture are enhanced by MR images obtained at different hospitals, since a common set of

attributes was identified as important in all six pairwise-testing tests. Such attributes include *sum of squares, sum average* and *difference entropy*. However, the particular variation of distance and direction of analysis varied across the six tests and was heavily reliant on the test-beds used, even when features were extracted from the same centre. For instance, features obtained from the first centre (BCH) showed an interesting pattern, where the ones that achieved optimal performance on the two test-beds were completely different. By inspecting Table 6.4, one can see that optimal feature-sets obtained from Tests 1 and 2 did not have any mutual features that were measured across the same direction and distance.

The aforementioned pattern, however, was not consistent with the results obtained with features acquired from the second centre (NUH). For instance, there were 23 mutual optimal features identified in Tests 3 and 4. The mutual optimal features identified in Tests 3 and 4 are mostly T2-weighted and are based on the GLCM technique: *Contrast, Entropy, Sum Entropy, Difference Entropy, Sum of Squares, Sum Average, Angular Second Moment and Inverse Difference Moment.* Variations of these features concerning inter-pixel distances and directions were present. The most frequent inter-pixel distance present in the mutual feature set was 4 pixels.

With regards to the third centre (GOSH), 8 of the features were identified to be optimal when tested on both test-beds (tests 5 and 6). These were mostly T2-weighted and all based on the GLCM technique: *(0,0,1) Correlation, (2,2,0) and (2,-2,0) Inverse Difference Moment, (0,0,2) Sum Average, (0,1,0) Sum Variance, (0,0,2) Sum Of Squares, (0,0,1)Difference Entropy, and 3D Kurtosis.*

The above observations suggest that whilst TA is, in principle, a scalable technique that can be used to classify tumour patterns using data gathered from other centres, there does not seem to be enough indication of any 'universal' features

that could be measured across specific directions and distances of analysis for use across centres, without taking other factors into account. Whilst there was a considerable number of attributes that were common across all six tests, none of the exact variation of features identified as optimal were mutual for all tests. The dependency of optimal performance on both acquisition centres and test-beds suggests that for TA to be used in practice, there needs to be a robust means of selecting the features for classifier training, which is likely to vary depending on each individual scenario. One potential solution is to combine the attributes that were identified as important across tests into a single score, perhaps through averaging, as a means of decreasing any inherent noise, increasing robustness and improving reliability. Additionally, meta-analysis of the performance of different feature selection methods need to drive future efforts in this area.

In terms of what had been reported in the literature, the closest work to this experiment is the multicentre study by Tantisatirapong et al [38], where conventional 2D TA was used in a binary classification problem to diagnose paediatric MB and PA, yielding an overall classification accuracy of 77% using T2-weighted data. Whilst it is not possible to directly compare the findings of this experiment to the current state-of-the-art, due to variations in primary aims and methodologies, the overall classification accuracies of 72% (before SMOTE) and 77% (after SMOTE) are in line with what had been reported in [38].

Although 3D TA of MRI relies heavily on sophisticated mathematical procedures, this study was entirely carried out using commercially available and open-source software, which are provided with well documented manuals to support their use by personnel with limited programming backgrounds.

6.5 Study Limitations and Future Work

The study discussed in this chapter suffers from the limitation that the presented pairwise testing results are a best case scenario, as the comparison looked into *optimal* AUC values. To determine robustness of the obtained findings, it will be necessary to further test the classifiers with the optimal settings identified in this study. This can be done using a three-fold validation approach, where training is done on one dataset, followed by a testing stage on another dataset where the optimal classifier settings are identified, and finally a validation stage where the identified optimal settings are tested for robustness.

Although the analysis was carried out on the three most frequently occurring paediatric brain tumours (MB, PA and EP), this methodology can be extended to other brain tumour types, provided enough data samples are available for use as a test-bed. It will also be interesting to look into the use of 3D TA on diffusion-weighted imaging (DWI), as work currently available in the literature has shown promising results with 2D TA of DWI.

6.6 Conclusions

In conclusion, the results of the study presented in this chapter indicated that despite the differences in textural information among MR images from different hospitals, feature-sets from one hospital may be used for successful tumour type classification when tested on data from other hospitals; an important finding for future clinical adoption of TA. The findings of the study presented here support the use of 3D TA on conventional MR images to aid diagnostic classification of paediatric brain tumours.

Chapter 7

Predicting Survival in Paediatric Medulloblastoma

Some aspects of the work presented here were published in [**P06**]. Publication details can be found on page *xx*.

7.1 Introduction

Although brain tumour characterisation using TA of MR images has received a great deal of attention over the past decade, much of this work concentrated on the diagnosis of tumour types and on the comparison of different algorithms, in order to establish which combinations yield the best performances. The encouraging results reported in Chapters 5 and 6 and in the literature, with regards to the efficacy of MRI TA, raise an interesting question: *If textural features could capture powerful patterns that aid the diagnosis of childhood brain tumours, can they also be used to predict patients' survival prognosis?*

Following diagnosis, determination of prognosis is an important step in brain tumour management, with implications that determine treatment options. Therefore, accurate non-invasive predictors of prognosis have the potential to advance clinical management of patients for therapy and the possibility to support more informed discussions with the patient's family.

To the best of the author's knowledge at the time of writing, there has been no published work on investigating brain tumour survival predictors based on image analysis of conventional MRI, such as T1 and T2-weighted scans. Such scans are routinely acquired when a patient is presented with a suspected brain tumour, and their reported success in diagnostic TA applications suggests a possibility that valuable but complex prognostic patterns may exist undiscovered in the data.

In this regard, the primary aim of the study presented in this chapter was to determine whether textural features extracted from conventional MR images were able to predict the survival of paediatric medulloblastoma: the most common

malignant brain tumour occurring in childhood. This was done by carrying out 3D TA on pre-contrast T1 and T2-weighted images using a number of different statistical TA techniques: histogram, absolute gradient, grey-level co-occurrence matrix (GLCM) and grey-level run-length matrix (GLRLM).

7.2 Materials and Methods

7.2.1 Cohort Details and Image Acquisition

The clinical material used in this retrospective study consisted of pre-contrast T1 and T2-weighted MR images of thirty-two children attending treatment at Birmingham Children's Hospital and subsequently diagnosed with medulloblastoma. All images were anonymised and obtained from a secure e-repository provided by CCLG [4]. The same acquistion protocols that were introduced in the single centre study (Chapter 5) were used to obtain the images.

Out of the thirty-two patients, one did not have T1-weighted data and five did not have T2-weighted data available on the e-repository, but they were still included in the study. Approval for the study was obtained from the research ethics committee, and informed consent was taken from guardians. In order to obtain diagnoses in accordance with the WHO classification, tumour samples were taken from all patients and underwent histopathological examinations.

7.2.2 Image Pre-processing

Similar to the studies presented in chapter 5 and 6, image pre-processing was carried out by manually selecting axial slices from the dataset, segmenting the tumour using the Snake GVF algorithm, and normalising the images using the $\mu + / - 3\sigma$ technique.

7.2.3 Extraction of Textural Features

MaZda software was used to carry out 3D TA based on the histogram statistics, absolute gradient, GLCM and GLRLM techniques, extracting the same features that were used in chapters 5 and 6.

7.2.4 Identifying Textural Features with Potential Prognostic Value

Including both imaging modalities and all four TA techniques in the analysis would result in a very large number of 566 features. It was therefore sensible to first investigate a sub-set of features that were likely to capture survival prognosis patterns well, and then test the identified features using a suitable survival analysis technique, such as *Kaplan-Meier* estimator. A supervised learning experiment was therefore carried out, in order to identify potentially optimal features, as detailed below. Python's Orange machine learning library was used to carry out the work explained in this section.

Step 1. Categorising the Feature-Set

The aim of this step was to organise the feature-set in a way that separates data of patients with good prognosis from those with poor prognosis. This is, however, complicated by two problems. Firstly, it is difficult to define what good prognosis is, in the sense that there is no particular cut-off value in terms of survival time, after which the patient is defined to have good survival. Secondly, assuming that we define a cut-off value, for example 4 years, the analysis would be further complicated by the fact that some patients are currently alive but their diagnosis date was less than 4 years ago. If these patients were to be included in the analysis, it would not be reasonable to categorise them as having good prognosis; because even

though they are currently alive, the cut-off point has not been reached. Similarly, including them under the poor prognosis category would have been problematic, because whether they will survive till at least the cut-off point remains unknown at the time of analysis.

In order to address the aforementioned issues, a cut-off value of 4 years was chosen, and if patients had survived until at least this point, they were categorised as having good prognosis. We proceeded by temporarily removing data of the 10 patients who are still alive but have not reached the cut-off point, which reduced our cohort size to 22 patients. The small cohort was well spread-out, where 9 patients had survived for at least 4 years and 13 patients had died before the 4 years point, implying that 4 years was a fairly suitable choice of cut-off value.

Step 2. Identifying Optimal Features Using a Supervised Learning Approach

The aim of this step was to apply supervised learning techniques to the categorised feature-set from Step 1, in order to identify a number of textural features that were likely to have strong prognostic value. Entropy-MDL discretisation was used to partition the textural features to a discrete number of intervals.The discretised sub-set was then used to train a simple Naive Bayes classifier in a two-class problem, with the aim of classifying the data points as *'had died before 4 years'* or *'had survived for at least 4 years'*. The purpose of doing this was to test whether the selected features did in fact capture discriminative prognostic patterns that enabled good classification rates across the two classes. Validation was carried out by randomly sampling the data points into a 50% training set and a 50% testing set. This process was repeated 5 times prior to calculating average classification rates. Encouraging classification results were obtained by the Bayesian classifier,

as detailed in the results section, which allowed us to proceed by testing individual features using a Kaplan-Meier estimator.

7.2.5 Statistical Methods

After a sub-set of potentially valuable textural features has been identified, it was necessary to examine its significance in a structured manner. *Kaplan-Meier survival estimator* was used to individually test each feature identified in the previous step, by examining whether a high feature value is associated with significant differences in survival time across the entire cohort of 32 patients. In order to establish whether a particular feature value was high or low, we used the cut-off values determined by Entropy-MDL discretisation algorithm during the feature selection step. Since these values lead to the training of an effective survival classification model, it was assumed that they were suitable choices as cut-offs.

In order to test the study's primary hypothesis, the log-rank test was used with a chosen p value of 0.05. Both *Kaplan-Meier* and *log-rank* tests were performed on MATLAB (version R2014b), using the KMPlot [68] and Logrank [67] libraries available from MATLAB File Exchange. Following this, *Pearson's pairwise linear correlation test* was applied, in order to identify any inherent links between the optimal features.

7.3 Results

The classification test used to identify potentially optimal features yielded a classification accuracy of 89% and an AUC value of 92%. Such results allowed us to proceed by testing individual features using a Kaplan-Meier estimator. Out of the 566 available features, only 36 were initially identified by Entropy-MDL to be of

Table 7.1: Summary of the textural features identified to be of significant prognostic value by log-rank test. Note that all features were extracted from T2-weighted images.

Feature	p
Sum Variance (1,-1,0)	0.00635 (<0.01)
Sum Variance (1,0,0)	0.01070 (<0.05)
Sum Variance (2,-2,0)	0.00374 (<0.01)
Sum Variance (2,0,0)	0.00238 (<0.01)
Sum Variance (0,0,3)	0.00006 (<0.01)
Sum of Squares (1,1,0)	0.00462 (<0.01)
Sum of Squares (0,0,3)	0.01565 (<0.05)
Angular Second Moment (0,2,0)	0.02099 (<0.05)
Angular Second Moment (2,2,0)	0.00374 (<0.01)
Angular Second Moment (2,-2,0)	0.00374 (<0.01)
Angular Second Moment (3,0,0)	0.00628 (<0.01)
Angular Second Moment (0,3,0)	0.00734 (<0.01)
Angular Second Moment (3,3,0)	0.00374 (<0.01)
Angular Second Moment (4,0,0)	0.00016 (<0.01)
Angular Second Moment (4,4,0)	0.00006 (<0.01)

potential value. Upon carrying out Kaplan-Meier and log-rank analyses, 15 of the 36 features were identified to be significant. A summary of the significant features, together with their associated p values, is shown in Table 7.1. An interesting observation is that all the features that we identified as significant were extracted from T2-weighted images.

Corresponding Kaplan-Meier survival plots for the 15 features are shown in Figures 7.1 to 7.4. Note that correlation coefficients, obtained from Pearson's pairwise linear correlation test, are available in Table 7.2.

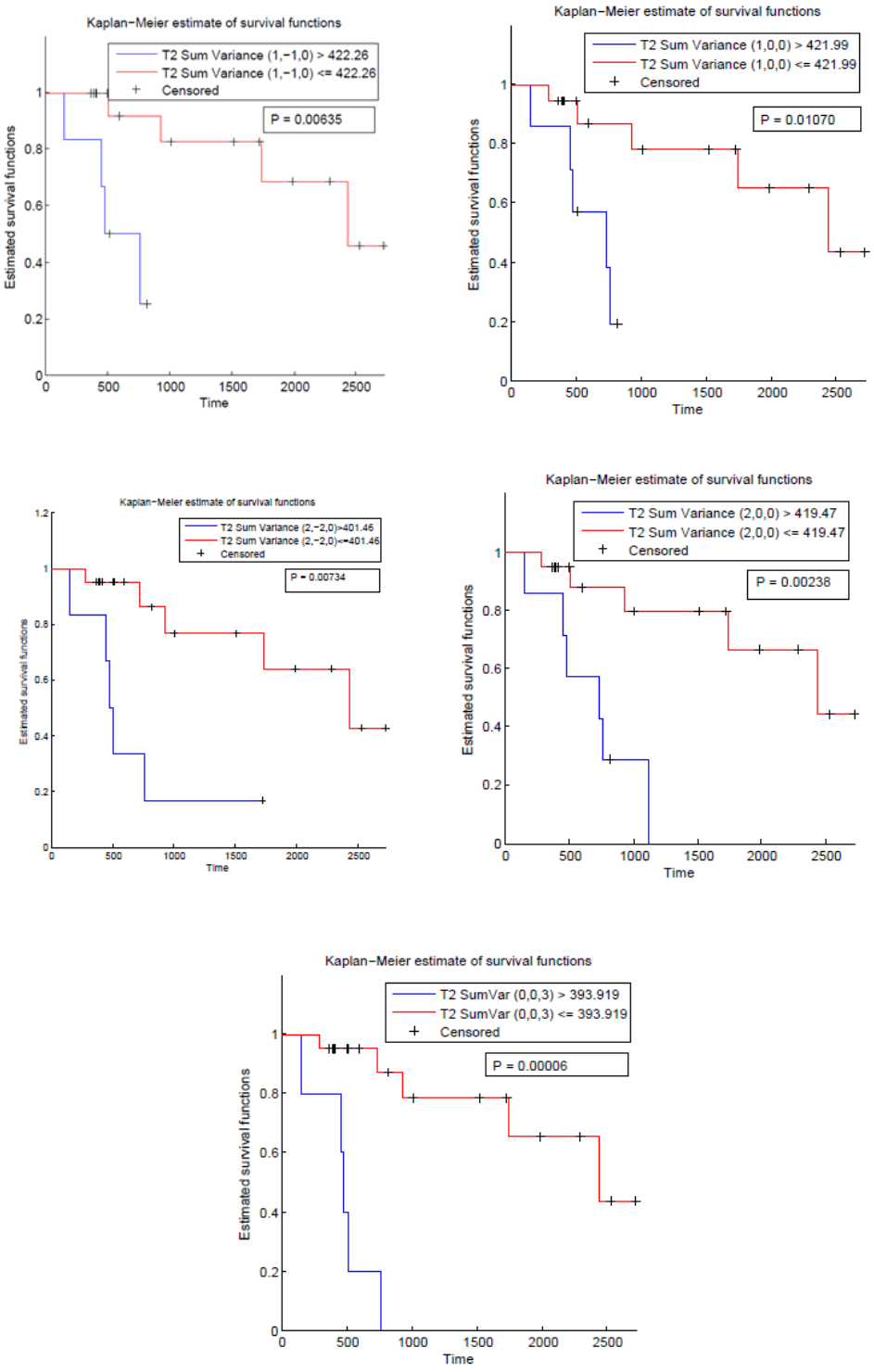


Figure 7.1: Kaplan-Meier survival curves for five of the fifteen features identified to be of prognostic value: *T2 Sum Variance* (1,-1,0),(1,0,0), (2,-2,0),(2,0,0),(0,0,3).

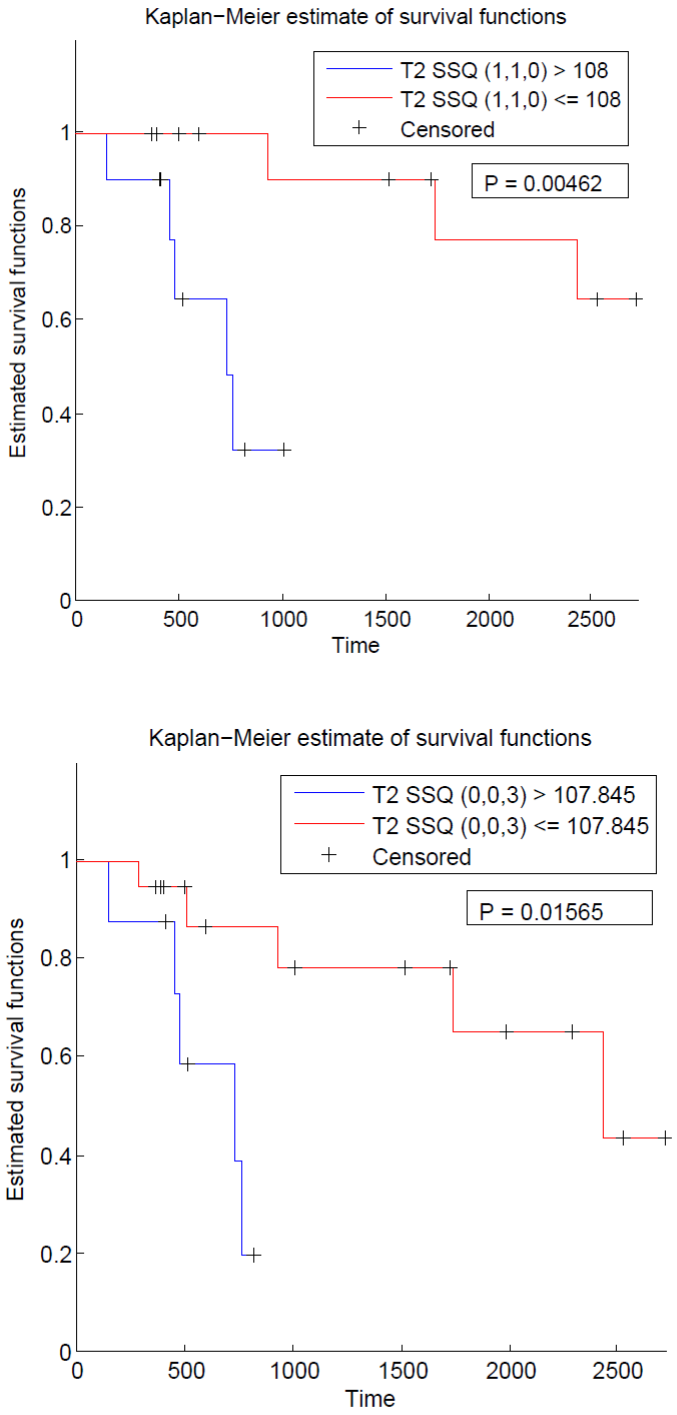


Figure 7.2: Kaplan-Meier survival curves for two of the fifteen features identified to be of prognostic value: *T2 Sum of Squares* (1,1,0) ,(0,0,3).

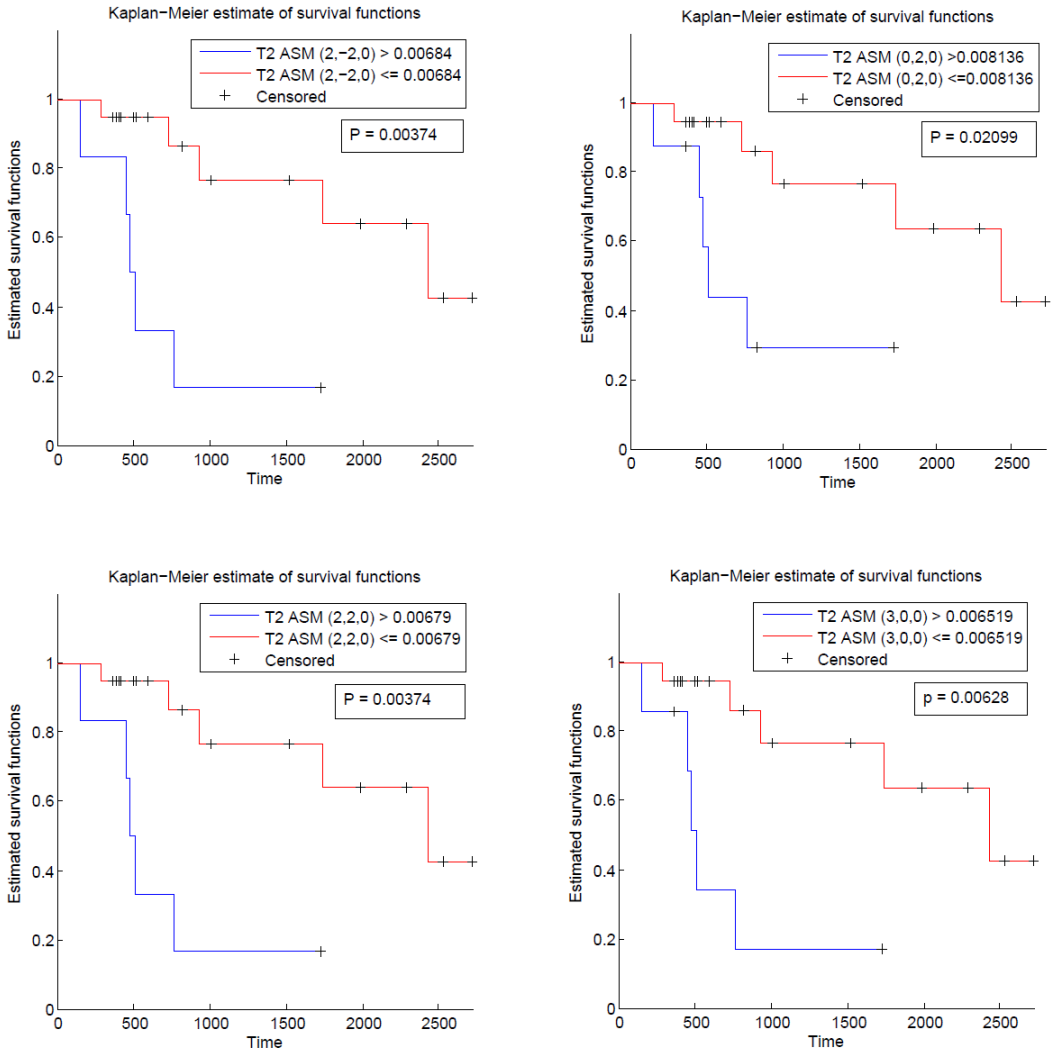


Figure 7.3: Kaplan-Meier survival curves for four of the fifteen features identified to be of prognostic value: *T2 Angular Second Moment* (2,-2,0), (0,2,0), (2,2,0), (3,0,0).

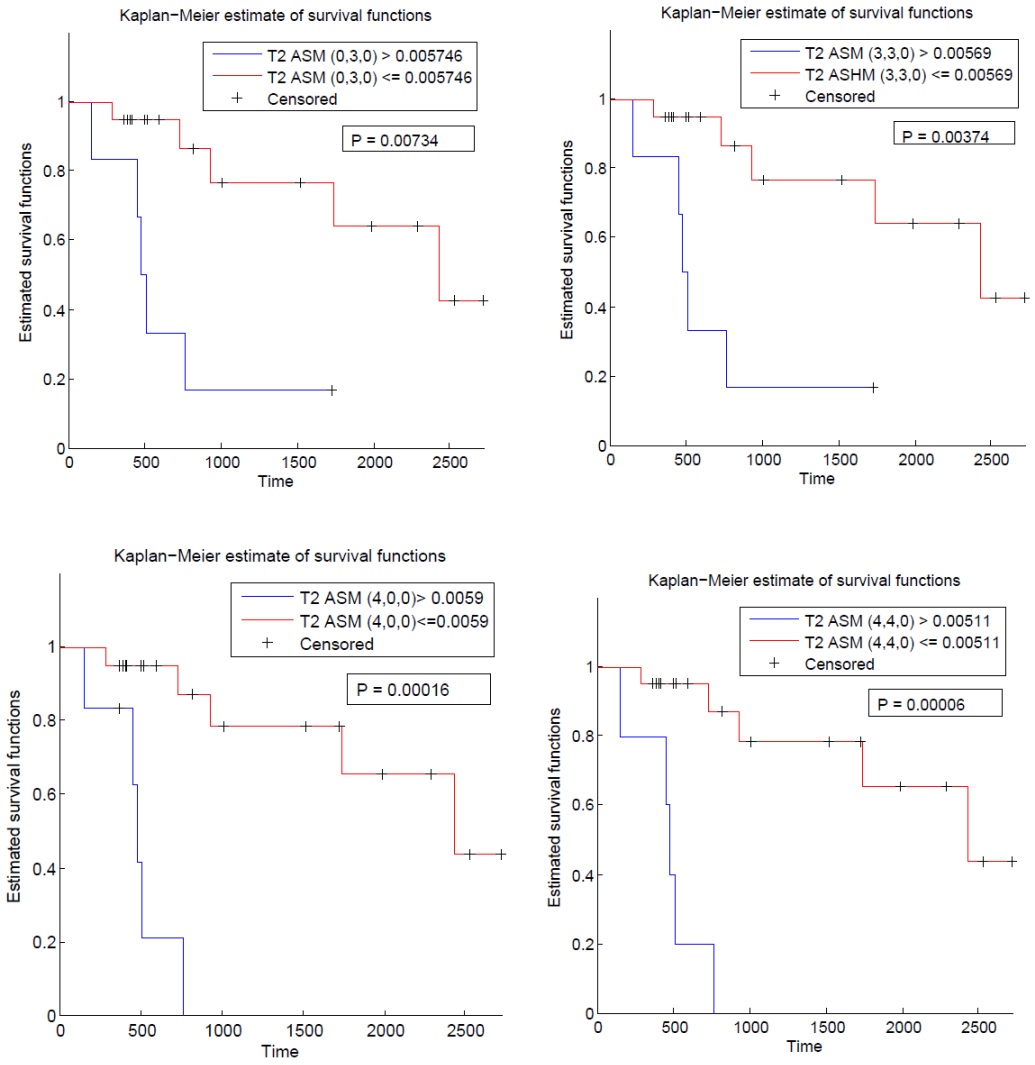


Figure 7.4: Kaplan-Meier survival curves for four of the fifteen features identified to be of prognostic value: *T2-weighted Angular Second Moment*(0,3,0),(3,3,0),(4,0,0),(4,4,0).

Table 7.2: A table summarising linear correlation coefficients obtained by applying Pearson's test on the 15 optimal features.

	SumVar	SumVar	SumVar	SumVar	SumVar	SSQ	SSQ	ASM	ASM	ASM	ASM	ASM	ASM	ASM
	(1,-1,0)	(1,0,0)	(2,-2,0)	(2,0,0)	(0,0,3)	(1,1,0)	(0,0,3)	(0,2,0)	(2,2,0)	(2,-2,0)	(3,0,0)	(0,3,0)	(3,3,0)	(4,0,0)	(4,4,0)
SumVar (1,-1,0)	1	0.9994	0.9983	0.9997	0.9956	0.9986	0.9987	0.614	0.5887	0.591	0.5491	0.553	0.5568	0.5749	0.5506
SumVar(1,0,0)	0.9994	1	0.9959	0.9988	0.9933	0.9997	0.9996	0.5979	0.5743	0.5765	0.536	0.5396	0.5432	0.5593	0.5372
SumVar (2,-2,0)	0.9983	0.9959	1	0.9988	0.9985	0.9943	0.9953	0.6397	0.6134	0.6161	0.5724	0.5772	0.5815	0.6015	0.5754
SumVar(2,0,0)	0.9997	0.9988	0.9988	1	0.9974	0.9976	0.9983	0.6173	0.5941	0.5964	0.5548	0.5589	0.5627	0.579	0.557
SumVar (0,0,3)	0.9956	0.9933	0.9985	0.9974	1	0.991	0.9933	0.6425	0.6235	0.6259	0.5845	0.5895	0.5936	0.6072	0.5892
SSQ(1,1,0)	0.9986	0.9997	0.9943	0.9976	0.991	1	0.9996	0.5948	0.5717	0.5738	0.5342	0.5376	0.5411	0.5566	0.5349
SSQ(0,0,3)	0.9987	0.9996	0.9953	0.9983	0.9933	0.9996	1	0.6007	0.5803	0.5823	0.5436	0.547	0.5505	0.5636	0.5446
ASM(0,2,0)	0.614	0.5979	0.6397	0.6173	0.6425	0.5948	0.6007	1	0.9661	0.9674	0.9284	0.9413	0.943	0.9947	0.9406
ASM(2,2,0)	0.5887	0.5743	0.6134	0.5941	0.6235	0.5717	0.5803	0.9661	1	0.9994	0.9826	0.9937	0.9937	0.9695	0.9938
ASM(2,-2,0)	0.591	0.5765	0.6161	0.5964	0.6259	0.5738	0.5823	0.9674	0.9994	1	0.9816	0.993	0.994	0.9705	0.9936
ASM(3,0,0)	0.5491	0.536	0.5724	0.5548	0.5845	0.5342	0.5436	0.9284	0.9826	0.9816	1	0.9807	0.9801	0.9365	0.9848
ASM(0,3,0)	0.553	0.5396	0.5772	0.5589	0.5895	0.5376	0.547	0.9413	0.9937	0.993	0.9807	1	0.9995	0.9526	0.9968
ASM(3,3,0)	0.5568	0.5432	0.5815	0.5627	0.5936	0.5411	0.5505	0.943	0.9937	0.994	0.9801	0.9995	1	0.9543	0.9972
ASM(4,0,0)	0.5749	0.5593	0.6015	0.579	0.6072	0.5566	0.5636	0.9947	0.9695	0.9705	0.9365	0.9526	0.9543	1	0.9535
ASM(4,4,0)	0.5506	0.5372	0.5754	0.557	0.5892	0.5349	0.5446	0.9406	0.9938	0.9936	0.9848	0.9968	0.9972	0.9535	1

7.4 Discussion

The results of this study have shown that 3D textural features, obtained non-invasively by analysing T1 and T2-weighted MR images, predict survival in a cohort of children diagnosed with medulloblastoma. Out of the initial feature-set that comprised 566 textural features of the statistical type, we were able to identify 15 features that hold useful prognostic value ($p < 0.05$). The 15 identified features are GLCM-based and are variations of sum variance, sum of squares and angular second moment, with different inter-pixel distances and directions of analysis.

An interesting finding was that variations of the same feature showed strong positive correlations between each other. For example, variations of Sum Variance: (1, -1,0), (1,0,0), (2, -2,0), (2,0,0), (0,0,3) had correlation coefficients above 0.9, as shown in Table 7.2. This suggests that carrying out the analyses in these particular combinations of directions and pixel distances could perhaps be measuring the same underlying pattern. Since one barrier for applying TA methods in clinical applications is the overabundance of techniques available for use, identifying such inherent relationships between features is an important step for simplicity in long-term clinical adoption.

Another interesting observation was the positive correlation between different features. For instance, Sum of Squares (SSQ) and Angular Second Moment (ASM) showed correlation coefficients that ranged between 0.53 and 0.60. This is interesting because SSQ is a measure of textural heterogeneity as it represents the spread around the central tendency; therefore it increases with the grey-level values spreading away from the mean. However, ASM is a measure of how constant or periodic the grey-level distribution is [20], [70]. A negative correlation would therefore be expected between the two features. Studying the Kaplan-Meier plots, which clearly show that high values of both features are associated with poor

survival prognosis, can also reflect the aforementioned positive correlation.

To explore this observation, ASM, SSQ and SumVar features were visualised on a number of tumour ROIs, as shown in Figure 7.5. This was done by generating feature maps on T2-weighted tumour ROIs, obtained by calculating textural features in a small window sliding over the image. The window was defined to be a 9x9 pixel mask, moving in steps of one pixel. Regions with high feature values appear brighter on the feature maps. By inspecting the maps on Figure 7.5, one could see how areas of relatively high ASM seem to be continuous tumour regions with similar grey-level intensities, whereas areas with high SSQ and SumVar seem to be edges where there is a sharp transition between grey-level values. Thus, it is likely that with increasing tumour complexities, there tends to be large a number of bulky regions, with similar grey-levels within each region (hence high ASM), and consequently more edges between them (hence high SSQ and SumVar). To aid with the illustration, the popular *peppers* image used in the image processing literature was included, together with its corresponding feature maps. As can be seen in the feature maps, there exists very high components of ASM, SSQ and SumVar in the same image.

7.5 Study Limitations and Future Work

Many children with brain tumours are treated on sophisticated protocols that stratify patients on an increasingly complex set of prgnostic markers that combine clinical information, conventional imaging, histopathological markers and tumour biology [69]. Thus, the prognostic role of TA within such protocols needs to be determined by its inclusion in large multi-centre trials in the future. Recent research efforts in the paediatric literature have identified glutamate as a biomarker for paediatric medulloblastoma [69]. Hence, identifying any inherent links between

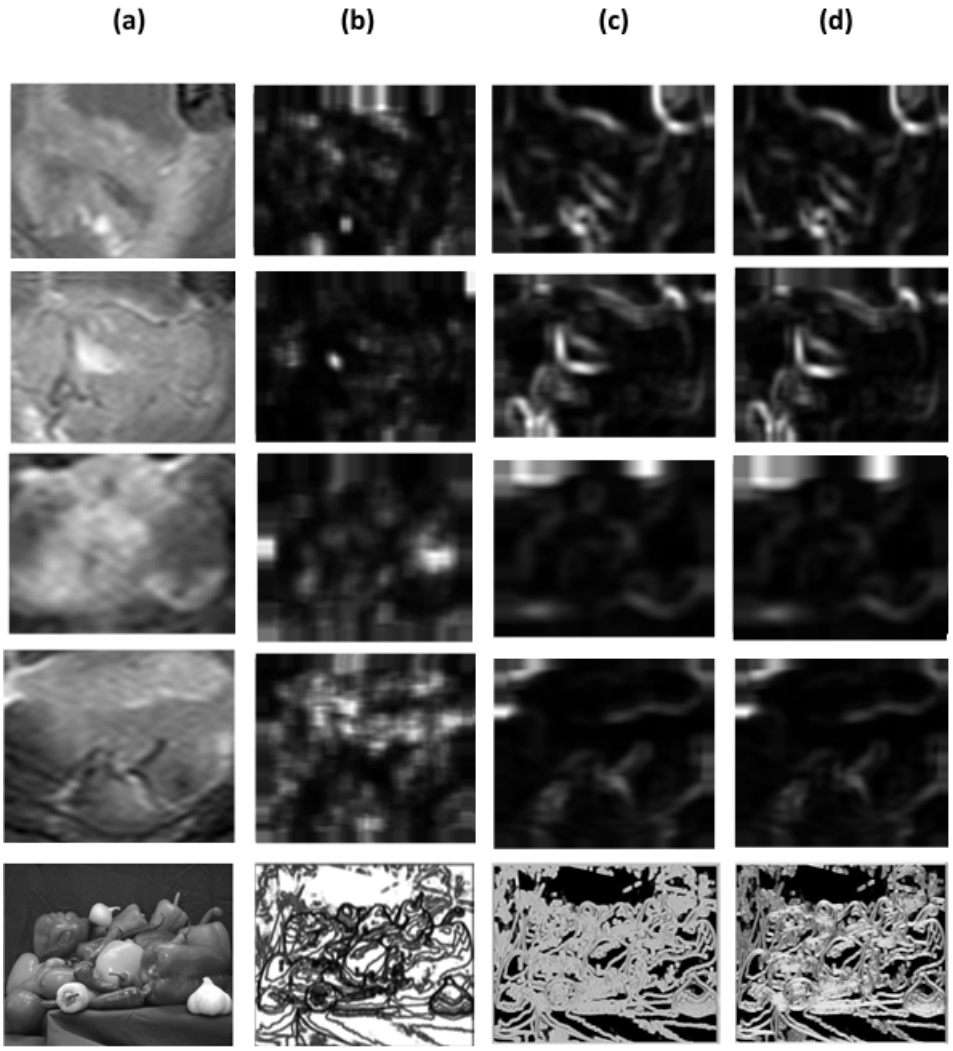


Figure 7.5: Four T2-weighted medulloblastoma ROIs and their corresponding feature maps based on (b) angular second moment (ASM) (c) sum of squares (SSQ) and (d) sum variance (SumVar). One could see how areas of relatively high ASM seem to be continuous tumour regions with similar grey-level intensities, whereas areas with high SSQ and SumVar seem to be edges where there is a sharp transition between grey-level values. To aid with the illustration, the popular *peppers* image used in the image processing literature was included, together with its corresponding feature maps. Original MR images were obtained from the CCLG database [4].

texture features and glutamate is another important future task.

An additional limitation to the present study is the relatively small number of medulloblastoma patients attending a single centre. Larger multicentre studies, similar to the investigation discussed in Chapter 6, are required to confirm the robustness of TA as a prognostic biomarker.

7.6 Conclusion

The third and final objective of this thesis, to investigate the prognostic value of MRI TA, was met in this chapter. The features that were identified as significant prognostic biomarkers for paediatric medulloblastoma were all T2-weighted and GLCM-based. Following diagnosis, determination of prognosis is an important step in tumour management, with implications that determine treatment options. Therefore, the identified features have the potential to advance clinical management of patients for therapy following cross-centre validation.

Chapter 8

Summary, Conclusion and Recommendations for Future Work

This chapter aims to discuss a summary of the main points arising from the research caried out as part of this thesis, and to identify the overall conclusions. A number of recommendations for future work are finally discussed.

8.1 Summary

The primary aim of this thesis was to explore the effectiveness of texture analysis of MR images in supporting clinicians with non-invasive characterisation of paediatric brain tumours. The work presented here contributed to the body of knowledge by addressing three major sub-problems that have received little attention in the paediatric literature: diagnostic classification, cross-centre transferability, and prognosis.

After identifying the primary aim and objectives of this thesis, it was essential to review the technique of MRI, in order to identify the underlying physical processes and to discuss how important MR imaging parameters, such as TE and TR, are linked to visualisation of common paediatric brain tumours. The importance of this was to understand the nature of the clinical data used in this thesis. This was achieved in Chapter 2.

Since much of this thesis was devoted to the use of textural features within a classification framework, it was necessary to review popular machine learning methods and algorithms, particularly classification models, feature selection algorithms, and ways of validating classifiers' performance. This was addressed in Chapter 3.

Thirdly, it was important to review popular texture analysis methods in order to identify the nature of radiomic information they can provide, and to thoroughly review the current state-of-the-art, particularly work available on brain cancer. In Chapter 4, existing statistical texture analysis methods (histogram, absolute

gradient, GLCM and GLRLM) were reviewed, and the nature of features they provide was discussed. With regards to the literature review, an important finding was the need for further research into maximising the value of textural features as diagnostic biomarkers, through the use of 3D analysis. Another important finding was the lack of studies that looked into the use of MRI textural features as prognostic biomarkers.

The thesis went on to advance the field of non-invasive tumour characterisation in paediatric neuro-oncology by exploring the effectiveness of textural features obtained from conventional MR images as diagnostic biomarkers. This was achieved through a single-centre study (Chapter 5). An important finding of the single-centre study was the importance of carrying out multi-slice (3D) TA to ensure that the diagnostic value of the technique is not diluted. Six machine learning classifiers were tested with 3D and 2D features, and the statistical significance of the obtained differences was rigorously analysed. It was also concluded that the choice of machine learning classifier is a less important question, since the differences in performance between different classification algorithms were not statistically significant. The experimental work presented in Chapter 5 met the first objective of this thesis.

The second objective of this thesis was to determine the diagnostic efficacy of TA on a multi-centric level. To this end, Chapter 6 introduced a classification study that used datasets obtained from three different hospitals and using scanners made by different manufacturers. With SVM yielding classification performances of up to 85% AUC, the cross-centre transferability of TA was shown possible. One issue that is commonly faced when designing such experiments is the issue of class imbalance, where tumour types are not equally represented in the cohort. We illustrated how the use of synthetic ependymoma samples can lead to a noticeable

increase in sensitivity (11% to 87%), and we recommend the use of such over-sampling techniques in future studies.

The final objective of this work was to determine whether textural features could potentially be used as prognostic biomarkers. This is important because following diagnosis, the determination of prognosis is an important step in brain tumour management, with implications that determine treatment options. The work presented in Chapter 7 met this objective through a survival analysis study, where fifteen features extracted from T2-weighted images were identified to be of significant prognostic value when tested on the cohort obtained from Birmingham Children's Hospital. This work was carried out on MR images of medulloblastoma; the most commonly occurring brain tumour in childhood, and the obtained success motivates further research into other tumour types.

8.2 Conclusion

The primary conclusion of this thesis is that MRI TA is valuable for the characterisation of paediatric brain tumours, providing quantitative information that can supplement visual inspections performed by radiologists. In recent years, it has been recognised that medical images contain more useful information than may be perceived with human vision, leading to the field of radiomics, whereby additional features can be extracted by computational techniques. Evidence has slowly accumulated showing that features obtained by TA of MR images can potentially provide a set of useful tools for non-invasive characterisation of paediatric brain tumours. In this thesis, the problem of tumour *characterisation* was divided into two parts: diagnosis and prognosis. In terms of specific findings, this work offers three novel contributions to knowledge. Firstly, it was shown, through experimental analysis, that TA can diagnostically classify common brain tumours

with high accuracies; and that such classification could be optimised by extending the analysis to include multi-slice features. The second contribution of this work was the analysis of the efficacy and cross-centre transferrability of TA, using multicentric, heterogeneous datasets. The findings suggested that TA is highly effective in diagnostic classification when tested on multicentre data. Thirdly, it was shown, through the analysis of clinical medulloblastoma data, that TA can be used to predict the survival prognosis of the patients, using features extracted from conventional T2-weighted images. On the basis of the findings of this thesis' experiments, it was shown that TA can potentially have a large clinical impact, since MR imaging is routinely used in the brain cancer clinical work-flow worldwide; providing an opportunity to improve decision-support at low cost.

8.3 Recommendations for Future Work

The work presented in Chapters 5 and 6 focused on diagnostic classification of MB, PA and EP using textural features of conventional MRI. We looked at the three most commonly occurring childhood brain tumours, which motivates the need to include rarer tumour types into future work. Besides being an important step towards clinical adoption of TA, including rarer types will ease comparisons with radiological performance, since the identification or ruling out of rarer tumours is likely to present radiologists with significant problems when trying to make a provisional diagnosis.

Moreover, it remains noteworthy that the variety of acquisition parameters used for the T1 and T2-weighted images may have affected the consistency of how tumours appeared across different scans, and consequently affected the ability of TA to characterise them. It is therefore possible that TA could perform better if the input is more uniform. With regards to the directional sensitivity of the

textural features used, the high correlation between them suggests that combining them into a single score, perhaps through averaging, might decrease noise and improve reliability. Exploring these two points is an important future extension to this research.

Another important future extension to this work is the inclusion of additional imaging types, such as post-contrast T1-weighted MRI. Although post-contrast MRI is routinely acquired at the centres from which the clinical data was obtained, it was not included in this study at this stage, due to a number of concerns with regards to standardisation. There are many variables that affect the images and make quantification difficult, particularly in children. For instance, the contrast bolus is injected by hand, sometimes into a peripheral vein whilst in others into a central line. Additionally, bolus duration is very variable, the time from injection to image acquisition is not standardised and cardiovascular parameters vary greatly. Visual inspection of these images bears out the differences in T1 post-contrast imaging between acquisitions. The method is useful for qualitative interpretation but further work is required before considering quantitative analysis.

In the author's opinion, the biggest limitation that MRI texture analysis studies presently suffer from is the lack of clear clinical meanings to the features identified as biomarkers. Establishing such meaning is a challenging task since TA, in theory, captures underlying MR imaging patterns that are below human vision. One way this issue could perhaps be tackled in the future is by carrying out TA on biopsy samples under different microscopic scales, where clinical attributes can be easily correlated with important textural features. Assuming that such meaning could be translated to MR imaging scales, this could potentially provide radiologists with a number of textural patterns to look for when carrying

out initial tumour characterisation. Good understanding of feature meanings will ensure that the generated knowledge and the explanation of classifier decisions will be *transparent* to the clinicians. This will support clinical acceptance of TA, since according to Kononenko [85], transparency is an important requirement for decision-support systems to be useful in solving medical diagnostic tasks.

The single-centre study discussed in Chapter 5 showed that whilst some classifiers might outperform others when tested on our cohort, such variations are not statistically significant, suggesting that the choice of which classifier to use is perhaps a less important question. Hence, an important factor to consider when choosing which classifier to apply in future research is the classifier's explanation ability. For instance, Naive Bayes and Classification Tree classifiers might be preferred by clinicians because of the nature of information they provide [85]. Alternatively, instead of selecting a single best classifier, combining their decisions when classifying a data point might be a more robust option.

Among the reasons for slow acceptance of decision support systems in clinical settings, perhaps the most reasonable one is that the introduction of such technologies will further increase the abundance of tools and instrumentation available to clinicians [85]. The use of non-invasive TA would have has the undesirable side effect of further increasing the complexity of the radiologist's work, which is already sufficiently complicated. Therefore, TA and machine learning systems will have to be integrated into the existing instrumentation that makes its adoption as natural as possible.

Appendix A

Preliminary Study

A.1 Introduction

The work presented here discusses a preliminary investigation that was conducted during the early stages of this research with the aim of answering the following question: *can MRI TA potentially be used, within a machine learning framework, to capture quantitative patterns for the characterisation of paediatric brain tumours?*

This preliminary study focused on two aspects: 1. the diagnostic potential of TA, and 2. potential cross-centre transferrability of the technique. These aspects were investigated by carrying out conventional 2D TA on multicentric MR imaging data. The datasets fell into the astrocytic, ependymal and embryonal histopathological categories. Due to the preliminary nature of this study, it was not included in the main body of this thesis. The positive findings of this study, however, motiavted rigorious analysis of MRI TA for tumour characterisation, as per Chapters 5 - 7.

Some aspects of the work presented here were published in [**P02**]. Publication details can be found on page *xx*.

A.2 Materials and Methods

A.2.1 Clinical Materials

Table A.1: A table summarising the datasets included in this preliminary study. Three histopathological tumour categories were included: astrocytic, ependymal and embryonal. All images were obtained from the CCLG database [4].

Astrocytic (26)	**Ependymal (19)**	**Embryonal (25)**
Pilocytic Astrocytoma (21)	Anaplastic Ependymoma (8)	Medulloblastoma (21)
Glioblastoma (5)	Ependymoma (11)	Atypical Teratoid0Rhadoid (4)

The dataset consisted of anonymised T2-weighted MR images of 70 children with verified and untreated brain tumours. All the chosen records fell into the astrocytic, ependymal or embryonal histopathological tumour categories. Image acquisition was carried out at three centres [1] using the following scanners: 1.5T Siemens Symphony, 1.5T Siemens Avanto (Siemens Healthcare, Erlangen, Germany), 1.5T General Electric Signa (GE Healthcare, Little Chalfont, UK), 1.5T Phillips Intera and 3T Phillips Achieva (Philips Healthcare, Amsterdam, Netherlands), following a common protocol defined by the Childrens Cancer and Leukaemia Group (CCLG) Functional Imaging Group.

All images were anonymised and held at a secure e-repository [4] provided by CCLG, from which the data was downloaded for use in this study. Although this experiment was a preliminary study, the decision to include datasets obtained from different hospitals was made in order to get an idea of the cross-centre transferrability of TA. Given the rarity of paediatric brain tumours, the included cohort of 70 patients was sufficiently large from a clinical perspective. In fact, the largest cohort size that was used in the paediatric brain tumour MRI TA literature included 50 patients, to the best of the author's knowledge at the time of writing [38]. Table A.1 shows the breakdown of the data used.

A.2.2 Image Pre-processing

The first pre-processing step was slice selection, where an axial slice containing the largest tumour region was manually chosen from each dataset, using RadiAnt DICOM viewer [61]. The selected axial slices were then imported to MaZda texture analysis software, which was developed by Materka et al [31]. To identify regions of interest (ROIs), equally sized square regions (30x30 pixels) were manually placed

[1] Birmingham Children's Hospital, Nottingham University Hospital and Great Ormond Street Hospital.

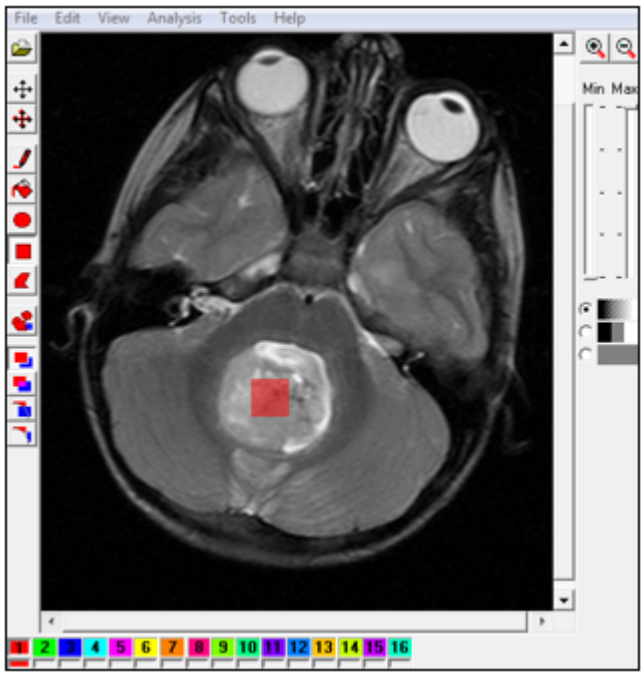


Figure A.1: A figure showing the placing of a 30x30 pixel region of interest on the tumour region of a T2-weighted image of a medulloblastoma (an embryonal tumour), using the MaZda software [31]. The original MR image was obtained from CCLG database [4].

inside the tumour areas. Since the data had been acquired from different imaging centres and using scanners produced by different manufacturers, there are usually variations in parameter settings that lead to the images having different grey-level ranges.

To mitigate the variations in parameter settings used while scanning different patients, the grey-level values within the identified ROIs were normalised through a two-step process (range selection and quantisation). Due to the preliminary nature of this study, grey-level range values were not manipulated. The second step involves quantising the resulting grey-level range between 1 to 2^k , where k is the number of bits per pixel. For instance, if our original range is between 1 and 1024, but we choose to use 8 bits per pixel, the dynamic range would be quantised to the range 1 to 256. In this study, 6 bits were chosen for quantisation.

A.2.3 Textural Features Extraction

After the images were pre-processed, each ROI was represented by 302 textural features computed using MaZda. The following techniques were used for feature extraction: histogram, absolute gradient, grey-level co-occurrence matrix (GLCM), grey-level run-length matrix (GLRLM), wavelets and autoregressive model. The statistical techniques were discussed in Chapter 4, and the non-statistical ones (wavelets and autoregressive model) are discussed in Appendix B. All the extracted features were based on the conventional 2D approach. In terms of GLCM and GLRLM, features were extracted for distances of 1, 2, 3 and 4 pixels in the horizontal, vertical and diagonal directions.

The feature sets that were computed from the ROIs were aggregated for analysis. The aggregated feature set was then re-organised into three separate groups, based on tumour histopathological category:

- Group (a): embryonal and astrocytic.
- Group (b): embryonal and ependymal.
- Group (c): astrocytic and ependymal.

A.2.4 Feature Selection and Supervised Learning

If all 302 features were evaluated together, it was very likely that the classification models would be over-fitted and poorly generalised. Irrelevant and redundant features are problematic because they may confuse the learning algorithm, by helping to obscure the distributions of the subset that holds influential features. The number of features tested therefore needed be reduced.

Orange [60], the python-based machine learning library (version 2.6a1) was used for feature selection. The entropy-MDL discretisation algorithm was em-

ployed to partition the features to a discrete number of intervals. As discussed in chapter 3, the entropy value of a feature can be a measure of its discriminative power, hence entropy-based discretisation can also be used for feature selection.

Supervised learning was then carried out, in a binary classification problem, on each of the three groups using the following algorithms:

- Naive Bayes (NB): Prior class probabilities based on: *Relative frequency*.
- K-Nearest Neighbour (kNN): Neighbours: *5*, Distance metric: *Euclidean*.
- Classification Tree (CTr): Attribute selection based on: *Information gain*.
- Support Vector Machines (SVM): Type: *C-SVM*, Kernel: *RBF*.

which were introduced in Chapter 3. The performance of the learning algorithms was evaluated using the random sampling strategy, with a relative training set size of 51%. The training/testing process was repeated 20 times to ensure realistic evaluation.

A.3 Preliminary Results

For each of the three groups, Table A.2 shows the classification accuracies obtained and their corresponding area under the receiver operator characteristics curves (AUC). Figures A.2-A.4 illustrate the ROC curves obtained with each group. The results suggested that textural features are potentially highly effective discriminants when comparing embryonal and astrocytic tumours, with kNN and NB achieving classification accuracies of 91%. The embryonal and ependymal group also yielded promising results, particularly with the Naive Bayes algorithm, which achieved a classification accuracy of 85%. Naive Bayes technique was similarly able to achieve the highest classification accuracy results for the astrocytic and ependymal group (74%).

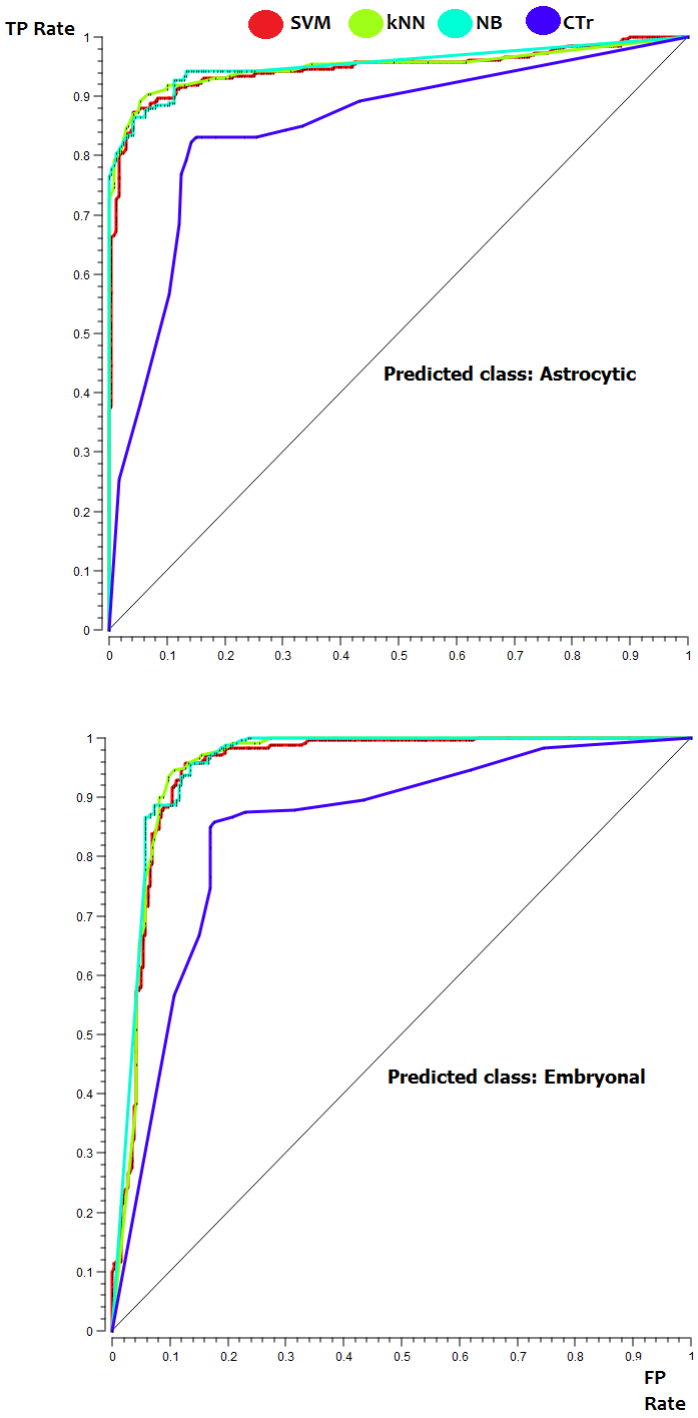


Figure A.2: A figure showing the receiver operator characteristics (ROC) curves for group (a): embryonal vs astrocytic.

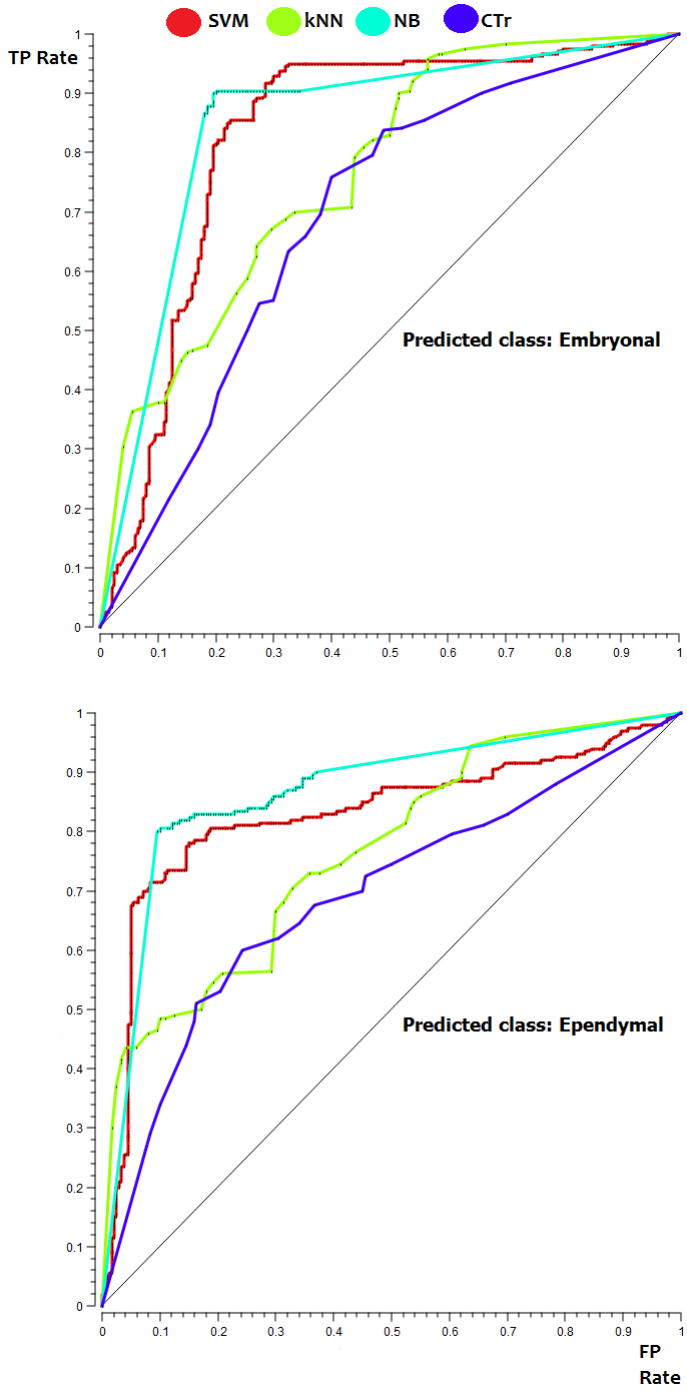


Figure A.3: A figure showing the receiver operator characteristics (ROC) curves for group (b): embryonal vs ependymal.

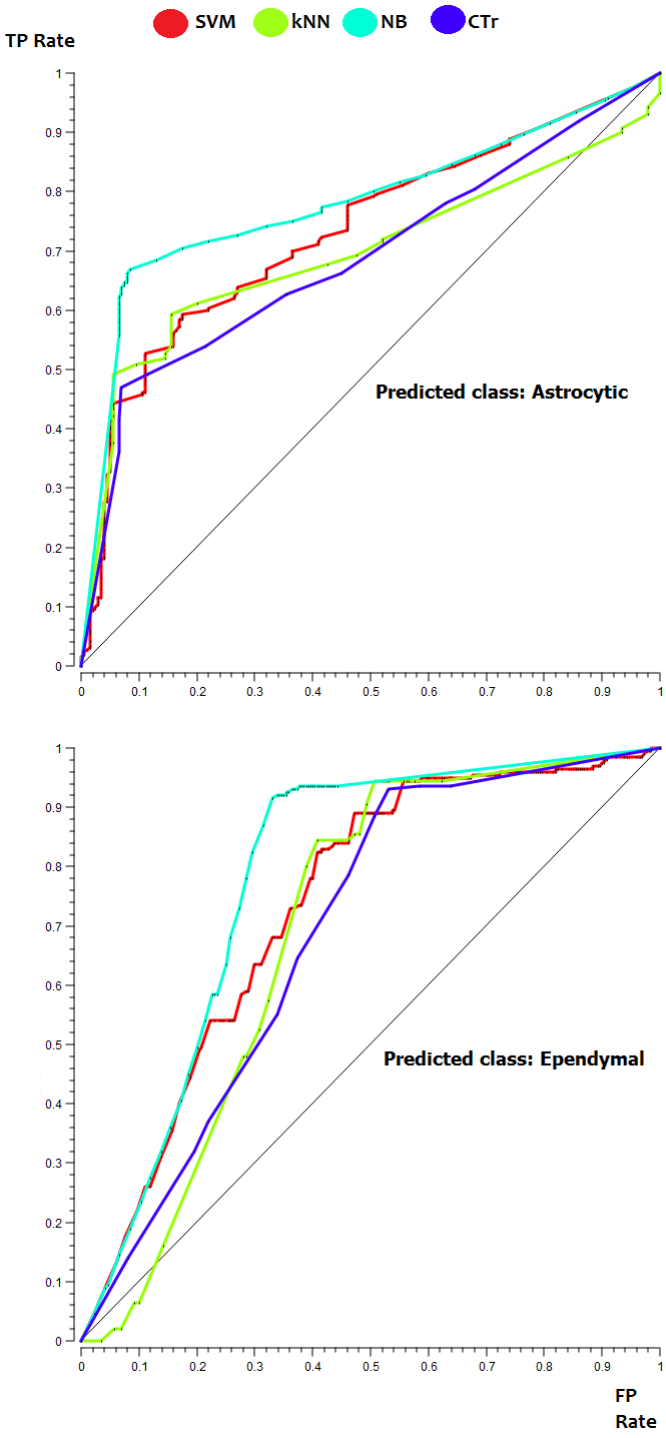


Figure A.4: A figure showing the receiver operator characteristics (ROC) curves for group (c): ependymal vs astrocytic.

Table A.2: A table summarising the classification accuracies and the corresponding AUC values obtained with each of the four classifiers on all three datasets. Model validation was carried out using random sampling, where the training/testing process was repeated 20 times.

Dataset	**Classification accuracy% (AUC%)**			
	SVM	kNN	NB	CTr
(a) Embryonal and Astrocytic	87(96)	91(96)	91(96)	84(84)
(b) Embryonal and Ependymal	82(88)	64(77)	85(88)	66(69)
(c) Astrocytic and Ependymal	68(76)	67(70)	74(80)	67(70)

A.4 Discussion

The findings were encouraging and motivated thorough experimentation into the efficacy of MRI texture analysis, within a machine learning framework, for diagnostic classification of childhood brain tumours. Additionally, the multicentric nature of the dataset used suggested that TA is potentially a scalable technique that could be used across multiple hospitals. Of course, the work described here was preliminary and was therefore succinct in its statistical analysis. However, a number of interesting points arised, some of which were mentioned in the Introduction section of this thesis. For instance: *Would using multiple MR slices and carrying 3D TA improve the classifiers' generalisation ability? Would the use of other modalities, such as T1-weighted images, add extra value to the classification? Can some of the included textural features be considered irrelevant or redundant, thus obscuring classifiers performance? Does the choice of learning algorithm significantly influence the classification results? And finally, is it feasible to characterise classifiers confidence when making diagnoses based on textural features, thus making it a practical tool for clinical settings?*

Additionally, the study suffered from a number of limitations, which were addressed in the experimental chapters of this thesis. For instance, the use of small-sized ROIs was certainly not sufficient for capturing enough information for

tumour discrimination. In practice, ROI outlining is usually performed manually, which is not only a time consuming task, but is also open to subjective interpretation of the radiologist. Therefore, a need to carry out future experiments using a robust segmentation technique that can capture ROIs that are representative of tumour patterns was identified. Additionally, the only performance metrics that were considered in this study were the classification accuracy and AUC measures, which on their own, might not be sufficient for thourough evaluation of classification performance. Finally, and perhaps most importantly, it is necessary to identify the nature of features that were selected during the dimensionality reduction stage in order to better understand the basis on which the learning models carried out their classification tasks.

The aforementioned points were addressed in the experimental parts of this thesis, particularly Chapter 5, which discusses a single-centre tumour classification study; the first main contribution of this thesis.

A.5 Conclusion

This study presented a preliminary investigation that looked into the classification of paediatric brain tumours into histopathological categories, using TA of T2-weighted MR images. This work was carried out at the early stages of this research with the aim of identifying whether textural patterns can potentially capture visual patterns that are beyond human vision, and subsequently be used for diagnostic classification in paediatric settings. It was found that, despite the use of limited textural patterns which were quantified using conventional 2D TA, it was possible to discriminate between tumour histopathological categoies in a binary classification problem. The encouraging preliminary findings motivated further research into maximising the technique's diagnostic value, which was addressed

in Chapter 5. Additionally, the multicentric nature of the data used encouraged rigurious analysis of TA's cross-centre transferrability, as discussed in Chapter 6.

Appendix B

Non-Statistical TA Techniques

Although the technical work presented in chapters 5-7 of this thesis is based on statistical TA techniques, two common non-statistical methods are introduced here as they had been used in a number of relevant studies in the literature, such as [37] and [38]. They were also used in the preliminary study presented in Appendix A.

(a) Autoregressive Model (Model-based Technique)

The *autoregressive (AR) model* is a linear prediction technique that aims to estimate future values of a signal as a linear function of previous samples. In TA, the AR modelling technique characterises statistical pixel dependencies by representing f_s, the grey-level intensity at location s, as a linear combination of surrounding grey-levels and an additive noise [81]. A causal AR model can be defined as:

$$f_s = \sum_{r \in N_s} \theta_r f_r + \epsilon_s \tag{B.1}$$

Where f_s is the image intensity at site s, ϵ_s is noise, N_s is a neighbourhood of s, and θ is a vector of AR model parameters. Features available from an AR model are:

- The coefficients for the four neighbouring pixels $(\theta_1, \theta_2, \theta_3, \theta_4)$. θ can be calculated as $\{\sum_s w_s w_s^T\}^{-1}\{\sum_s w_s f_s\}$, where $w_s = col[f_s, i \in N_s]$ [21].
- The standard deviation (σ) of the noise, where $\sigma^2 = R^{-1}\sum_s\{f_s - \theta w_s\}^2$. R is the number of pixels inside the ROI such that for the point s is moved to a pixel location, all the 4 immediate neighbours of s (Fig B.1) will be placed inside the ROI as well [21].

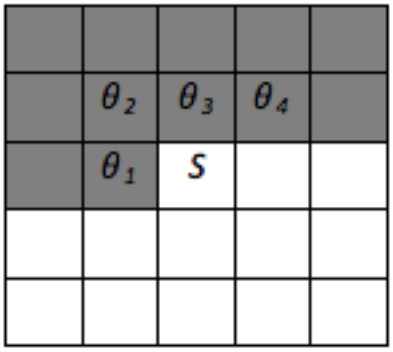


Figure B.1: A hypothetical pixel neighbourhood showing a pixel s and its surrounding region (shaded in grey) where a casual AR model neighbourhood may be located.

The AR modelling technique therefore assumes local interaction between local image pixels and can give an indication on how smooth or coarse the texture is. In other words, it is a way of describing shapes within an image by finding relations between groups of neighbouring pixels. The coefficient θ can be interpreted as a measure of statistical similarity between intensities of pixel s and its neighbours [21]. For coarse textures, the coefficients of neighbouring pixels will be widely varied, while in smooth ones the coefficients will be similar to each other [29]. Figure 4.9 shows an element s with the shaded area representing the region where a causal half-plane AR model neighbourhood may be located.

(b) Wavelet Analysis (Transform-based Technique)

In broad terms, *Wavelet transform* is a technique that can be used to separate data into different frequency components. The motivation behind the use of wavelets in TA is that features could be extracted at different imaging scales. To elaborate, consider the example of a constant MR scanner field of view, say 12.8 x 12.8 cm. If the slice thickness was changed from 1mm, through 0.5mm, to 0.25mm, one obtains images that contain 128x128 pixels, 256x256 pixels and 512x512 pixels respectively, leading to variations in textural pattern dimensions across these images [21]. Wavelet transform is performed through the use of a cascade of low (L) and high (H) pass filters. A single line or column of an image can be treated as a one-dimensional signal [21]. Wavelet transform is carried out on an image by first transforming all image rows, followed by all image columns [32], yielding four different sub-bands: *LL, LH, HL and HH.*

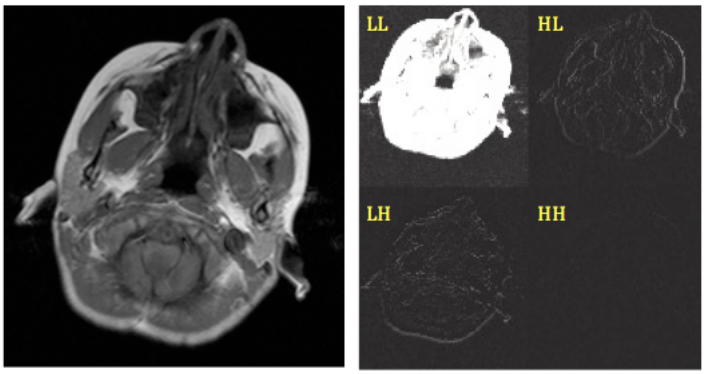


Figure B.2: An axial T2-weighted MR image (left) and its corresponding wavelet transform sub-bands (right). Original image was obtained from CCLG database [4].

The example brain MR image in Figure B.2 shows the general form of 2D wavelet transform; one can see that most of the images information is compacted in the LL sub-band. [1]

[1]Note that the brightness of the sub-bands has been manipulated to aid visualisation.

References

[1] Z.-P. Liang and P. C. Lauterbur, *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective,* SPIE Optical Engineering Press, 2000.

[2] R. Freeman, *Magnetic Resonance in Chemistry and Medicine,* Oxford University Press, 2003.

[3] E. Hahn, "Spin Echoes," *Physical Review,* vol. 80, no. 4, pp. 580-594, 1950.

[4] T. Arvanitis, K. Natarajan, J. Rossiter, J. Ting, Y. Sun, M. Wilson, N. Davies, E. Orphanidou-Vlachou, R. Grazier, J. Crouch, D. Auer, C. Clark, R. Grundy, D. Hargrave, F. Howe, T. Jaspan, M. Leach, L. MacPherson, G. Payne, D. Saunders, and A. Peet, "The childrens cancer and leukaemia group (CCLG) functional imaging e-repository for clinical trials of childhood brain tumours," *Neuro-Oncology*, vol. 12, no. 6, pp.118, 2010.

[5] Macmillan Cancer Support. *The brain - structure and function* [Online]. Available: `http://www.macmillan.org.uk/Cancerinformation/Cancertypes/Brain/Aboutbraintumours/Thebrain.aspx.`

[6] New York Presbyterian Hospital. *Posterior Fossa Tumours* [Online]. Available: `http://nyp.org/health/neuro-posterior-fossa.html`

[7] F. Davis and B. McCarthy, "Epidemiology of brain tumors," *Current Opinion in Neurology,* vol. 13, pp. 645 - 640, 2000.

[8] R. Rosebud, C. Lynch, J. Michael, and M. Hart, "Medulloblastoma: A Population-based Study of 532 Cases," *Journal of Neuropathology and Experimental Neurology*, 1991.

[9] K. K. Koeller and E. J. Rushing, "From the archives of the AFIP: medulloblastoma: a comprehensive review with radiologic-pathologic correlation," *Radiographics*, vol. 23, no. 6, pp. 1613 - 37, Jan. 2003.

[10] W. T. OBrien, "Imaging of Primary Posterior Fossa Brain Tumors in Children," *Journal of the American Osteopathic College of Radiology*, vol. 2, no. 3, 2013. .

[11] A. Poretti, A. Meoded, and T. A. G. . Huisman, "Neuroimaging of pediatric posterior fossa tumors including review of the literature," *Journal of Magnetic Resonance Imaging*, vol. 35, no. 1, pp. 32 - 47, Jan. 2012.

[12] T. P. Naidich and R. A. Zimmerman, "Primary brain tumors in children," *Seminars in Roentgenology,* vol. 19, no. 2, pp. 100 -114, Apr. 1984.

[13] K. K. Koeller and G. D. Sandberg, "From the archives of the AFIP. Cerebral intraventricular neoplasms: radiologic-pathologic correlation," *Radiographics,* vol. 22, no. 6, pp. 1473 - 505, Jan. 2002.

[14] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, B. W. Scheithauer, and P. Kleihues, "The 2007 WHO classification of tumours of the central nervous system," *Acta Neuropatholica,* vol. 114, no. 2, pp. 97 - 109, Aug. 2007.

[15] Flickr - Photo Sharing. *la Muzza* [Online]. Available: `https://www.flickr.com/photos/sergiotumm/15725948227/in/explore-2014-11-30/lightbox/`

[16] O. S. Al-Kadi, "Tumour Grading and Discrimination based on Class Assignment and Quantitative Texture Analysis Techniques." PhD Thesis, University of Sussex, 2009.

[17] Oxford Dictionary. *Texture - definition of Texture in English from the Oxford dictionary* [Online]. Available: `http://www.oxforddictionaries.com/definition/english/texture`

[18] W. H. Nailon, "Texture Analysis Methods for Medical Image Characterisation," Biomedical Imaging, Youxin Mao (Ed.), ISBN:978-953-307-701-1, In-Tech, 2010.

[19] M. Tuceryan and A. K. Jain, Texture Analysis, "Principles of Magnetic Resonance Imaging: A Signal Processing Perspective," *The Handbook of Pattern Recognition and Computer Vision*, P.S.P. Wang (Eds.), pp. 207-248, World Scientific Publishing, 1998.

[20] K. K. Holli, "Texture analysis as a tool for tissue characterization in clinical MRI." PhD Thesis, Tampere University of Technology, 2011.

[21] M. Hajek, "*Texture Analysis for Magnetic Resonance Imaging,"* European Network Cost Action B21, 2006.

[22] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 3, no. 6, pp. 610-621, Nov. 1973.

[23] L. Ma, T. Tan, Y. Wang, and D. Zhang, "Personal identification based on iris texture analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 25, no. 12, pp. 1519-1533, Dec. 2003.

[24] H. Yu, M. Li, H.-J. Zhang, and J. Feng, "Color texture moments for content-based image retrieval," *Proceedings of the International Conference on Image Processing,* vol. 1, pp. 929-932, 2002.

[25] E. I. Zacharaki, S. Wang, S. Chawla, D. Soo Yoo, R. Wolf, E. R. Melhem, and C. Davatzikos, "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme," *Magnetic Resonance in Medicine,* vol. 62, no. 6, pp. 1609-18, Dec. 2009.

[26] W. Chen, M. L. Giger, H. Li, U. Bick, and G. M. Newstead, "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images," *Magnetic Resonance in Medicine,* vol. 58, no. 3, pp. 562-571, Sep. 2007.

[27] L. C. V Harrison, M. Raunio, K. K. Holli, T. Luukkaala, S. Savio, I. Elovaara, S. Soimakallio, H. J. Eskola, and P. Dastidar, "MRI texture analysis in multiple sclerosis: toward a clinical analysis protocol," *Academic Radiology,* vol. 17, no. 6, pp. 696-707, Jun. 2010.

[28] G. Castellano, L. Bonilha, L. M. Li, and F. Cendes, "Texture analysis of medical images," *Clinical Radiology,* vol. 59, no. 12, pp. 1061-9, Dec. 2004.

[29] Technical University of Lodz, Institute of Electronics. MaZda User Manual [Online]. Available: `http://www.eletel.p.lodz.pl/mazda/download/mazda_manual.pdf`

[30] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing,* vol. 4, no. 2, pp. 172-179, 1975.

[31] A. M. and A. K. M. Strzelecki, P. Szczypinski, "A software tool for automatic classification and segmentation of 2D/3D medical images," *Nuclear Instruments and Methods in Physics Research,* vol. 702, pp. 137-140, 2013.

[32] M. Kociolek, A. Materka, M. Strzelecki, and P. Szcypinski, "Discrete wavelet transform derived features for digital image texture analysis," *Proceedings of the International Conference on Signals, Electronics and Systems,* pp. 163-168, 2001.

[33] P. Georgiadis, D. Cavouras, I. Kalatzis, D. Glotsos, E. Athanasiadis, S. Kostopoulos, K. Sifaki, M. Malamas, G. Nikiforidis, and E. Solomou, "Enhancing the discrimination accuracy between metastases, gliomas and meningiomas on brain MRI by volumetric textural features and ensemble pattern recognition methods," *Magnetic Resonance Imaging,* vol. 27, no. 1, pp. 120-130, Jan. 2009.

[34] D. Mahmoud-Ghoneim, G. Toussaint, J. M. Constans, and J. D. de Certaines, "Three dimensional texture analysis in MRI: a preliminary evaluation in gliomas," *Magnetic Resonance Imaging,* vol. 21, no. 9, pp. 983-987, Nov. 2003.

[35] F. Davnall, C. S. P. Yip, G. Ljungqvist, M. Selmi, F. Ng, B. Sanghera, B. Ganeshan, K. A. Miles, G. J. Cook, and V. Goh, "Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?," *Insights Imaging,* vol. 3, no. 6, pp. 573-89, Dec. 2012.

[36] D. Rodriguez Gutierrez, A. Awwad, L. Meijer, M. Manita, T. Jaspan, R. A. Dineen, R. G. Grundy, and D. P. Auer, "Metrics and Textural Features of MRI Diffusion to Improve Classification of Pediatric Posterior Fossa Tumors," *American Journal of Neuroradiology,* vol. 35, pp. 1009-1015, Dec. 2013.

[37] E. Orphanidou-Vlachou, N. Vlachos, N. P. Davies, T. N. Arvanitis, R. G. Grundy, and A. C. Peet "Texture analysis of T1 - and T2 -weighted MR images and use of probabilistic neural network to discriminate posterior fossa tumours in children," *NMR in Biomedicine,* vol. 27, pp. 632-639, Jun. 2014.

[38] S. Tantisatirapong, N. P. Davies, L. Abernethy, D. P. Auer, C. A. Clark, R. Grundy, T. Jaspan, D. Hargrave, L. MacPherson, M. O. Leach, G. S. Payne, B. L. Pizer, A. C. Peet, and T. N. Arvanitis, "Automated Processing Pipeline for Texture Analysis of Childhood Brain Tumours based on Multimodal Magnetic Resonance Imaging," *Biomedical Engineering*, vol. 791, pp. 791-081, 2013.

[39] R. A. Lerski, K. Straughan, L. R. Schad, D. Boyce, S. Blml, and I. Zuna, "MR image texture analysis - An approach to tissue characterization," *Magnetic Resonance Imaging*, vol. 11, no. 6, pp. 873-887, Jan. 1993.

[40] D. Glotsos, P. Spyridonos, D. Cavouras, P. Ravazoula, P. A. Dadioti, and G. Nikiforidis, "An image-analysis system based on support vector machines for automatic grade diagnosis of brain-tumour astrocytomas in clinical routine," *Medical Informatics and the Internet in Medicine*, vol. 30, no. 3, pp. 179-93, Sep. 2005.

[41] P. Georgiadis, D. Cavouras, I. Kalatzis, A. Daskalakis, G. C. Kagadis, K. Sifaki, M. Malamas, G. Nikiforidis, and E. Solomou, "Improving brain tumor characterization on MRI by probabilistic neural networks and non-linear transformation of textural features," *Computer Methods and Programs in Biomedicine*, vol. 89, no. 1, pp. 24-32, Jan. 2008.

[42] P. Georgiadis, S. Kostopoulos, D. Cavouras, D. Glotsos, I. Kalatzis, K. Sifaki, M. Malamas, E. Solomou, and G. Nikiforidis, "Quantitative combination of volumetric MR imaging and MR spectroscopy data for the discrimination of

meningiomas from metastatic brain tumors by means of pattern recognition," *Magnetic Resonance Imaging*, vol. 29, no. 4, pp. 525-35, May 2011.

[43] K. Holli, A.-L. Laaperi, L. Harrison, T. Luukkaala, T. Toivonen, P. Ryymin, P. Dastidar, S. Soimakallio, and H. Eskola, "Characterization of breast cancer types by texture analysis of magnetic resonance images," *Academic Radiology*, vol. 17, no. 2, pp. 135-41, Feb. 2010.

[44] D. Duda, M. Kretowski, R. Mathieu, R. de Crevoisier, and J. Bezy-Wendling, "Multi-Image Texture Analysis in Classification of Prostatic Tissues from MRI. Preliminary Results," *Information Technology in Biomedicine*, vol. 3, pp. 139-150, 2014.

[45] M. S. de Oliveira, M. L. F. Balthazar, A. DAbreu, C. L. Yasuda, B. P. Damasceno, F. Cendes, and G. Castellano, "MR imaging texture analysis of the corpus callosum and thalamus in amnestic mild cognitive impairment and mild Alzheimer disease," *American Journal of Neuroradiology*, vol. 32, no. 1, pp. 60-6, Jan. 2011.

[46] J. Zhang, L. Tong, L. Wang, and N. Li, "Texture analysis of multiple sclerosis: a comparative study," *Magnetic Resonance Imaging*, vol. 26, no. 8, pp. 1160-6, Oct. 2008.

[47] K. Jafari-Khouzani, K. Elisevich, S. Patel, B. Smith, and H. Soltanian-Zadeh, "FLAIR signal and texture analysis for lateralizing mesial temporal lobe epilepsy," *Neuroimage*, vol. 49, no. 2, pp. 1559-71, Jan. 2010.

[48] B. Ganeshan, E. Panayiotou, K. Burnand, S. Dizdarevic, and K. Miles, "Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival," *European Radiology*, vol. 22, no. 4, pp. 796-802, Apr. 2012.

[49] B. Ganeshan, K. Skogen, I. Pressney, D. Coutroubis, and K. Miles, "Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival," *Clinical Radiology*, vol. 67, no. 2, pp. 157-64, Feb. 2012.

[50] K. A. Miles, B. Ganeshan, M. R. Griffiths, R. C. D. Young, and C. R. Chatwin, "Colorectal cancer: texture analysis of portal phase hepatic CT images as a potential marker of survival," *Radiology*, vol. 250, no. 2, pp. 444-52, Feb. 2009.

[51] Y. S. Abu-Mostafa, M. Magdon-Ismail, H. Lin, *Learning from Data, a Short Course.* AMLbook, 2012.

[52] G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer Texts in Statistics, 2013 .

[53] Netflix Prize. *Netflix Prize* [Online]. `Available:http://www.netflixprize.com/`

[54] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[55] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Researchers," *ReCALL 31.,* 2004.

[56] I. Kononenko and E. Simec, "Induction of decision trees using ReliefF" *Proceedings of the ISSEK 94* no. 363, pp. 199220.

[57] U. M. Fayyad and K. B. Irani, "Multi-Interval Discretization of Continuos-Valued Attributes for Classification Learning," *Proceedings of the International Joint Conference on Uncertainty in AI* Q334 .I57, pp. 1022-1027, 1993.

[58] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis* vol. 1, no. 97, pp. 131-156, 1997.

[59] G. Koller, "Towards Optimal Feature Selection" *Tech. Report. Stanford infolab*

[60] J. Demar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik, and B. Zupan, "Orange: data mining toolbox in python," *Journal of Machine Learning Research,* vol. 14, no. 1, pp. 2349-2353, Jan. 2013.

[61] Medixant. RadiAnt DICOM Viewer [Online]. Available `http://www.radiantviewer.com/`

[62] P. Mansfield, "Imaging by nuclear magnetic resonance," *Journal of Physics E: Scientific Instruments* vol. 21, no.1, pp. 18, 1998

[63] G. Collewet, M. Strzelecki, and F. Mariette, "IInfluence of MRI acquisition protocols and image intensity normalization methods on texture classification," *Magnetic Resonance Imaging*, vol. 22, no. 1, pp. 81-91, Jan. 2004.

[64] S. L. Salzberg, "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 317-328.

[65] D. H. Jeong, C. Ziemkiewicz, W. Ribrasky, and R. Chang, Understanding Principal Component Analysis Using a Visual Analytics Tool, Charlotte Visualization Center, UNC Charlotte. [Online]. Available: `https://www.purdue.edu/discoverypark/vaccine/assets/pdfs/publications/pdf/UnderstandingPrincipalComponent.pdf`

[66] D. H. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang, "iPCA: An Interactive System for PCA-based Visual Analytics," *Computer Graphics Forum,* vol. 28, no. 3, pp. 767-774, Jun. 2009.

[67] G. Cardillo, Logrank - File Exchange - MATLAB Central [Online]. Available: `http://www.mathworks.co.uk/matlabcentral/fileexchange/22317-logrank.`

[68] G. Cardillo, KMPlot - File Exchange - MATLAB Central [Online]. Available: `http://www.mathworks.com/matlabcentral/fileexchange/22293-kmplot.`

[69] M. Wilson, S. K. Gill, L. MacPherson, M. English, T. N. Arvanitis, and A. C. Peet, "Non-invasive detection of glutamate predicts survival in pediatric medulloblastoma," *Clinical Cancer Research*, vol. 20, no. 17, pp. 4532-9, Jun. 2014.

[70] L. Harrison "Clinical Applicability of MRI Texture Analysis." PhD Thesis, University of Tampere, 2011.

[71] M. E. Mayerhoefer, M. J. Breitenseher, J. Kramer, N. Aigner, S. Hofmann, and A. Materka, "Texture analysis for tissue discrimination on T1-weighted MR images of the knee joint in a multicenter study: Transferability of texture features and comparison of feature selection methods and classifiers," *Journal of Magnetic Resonance Imaging*, vol. 22, no. 5, pp. 674-80, Nov. 2005.

[72] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Transactions on Image Processing,* vol. 7, no. 3, pp. 359-69, Jan. 1998.

[73] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[74] Cancer Research UK. Childhood Cancer Statistics [Online]. Available: `http://www.cancerresearchuk.org/cancer-info/cancerstats/childhoodcancer/`

[75] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, B. W. Scheithauer, P. Kleihues "The 2007 WHO Classification of Tumours of the Central Nervous System," *ACTA Neuropathologica,* vol. 1114, pp. 87 - 109.

[76] A. C. Peet, T. N. Arvanitis, D. P. Auer, N. P. Davies, D. Hargrave, F. a Howe, T. Jaspan, M. O. Leach, D. Macarthur, L. MacPherson, P. S. Morgan, K. Natarajan, G. S. Payne, D. Saunders, and R. G. Grundy, "The value of magnetic resonance spectroscopy in tumour imaging," *Archives of Disease in Childhood* vol. 93, no. 9, pp. 725-727, Sep. 2008.

[77] J. Vicente, E. Fuster-Garcia, S. Tortajada, J. M. Garca-Gmez, N. Davies, K. Natarajan, M. Wilson, R. G. Grundy, P. Wesseling, D. Monlen, B. Celda, M. Robles, and A. C. Peet, "Accurate classification of childhood brain tumours by in vivo 1H MRS - a multi-centre study," *European Journal of Cancer*, vol. 49, no. 3, pp. 658-67, Feb. 2013.

[78] A. Kassner and R. E. Thornhill, "Texture Analysis: A Review of Neurologic MR Imaging Applications," *American Journal of Neuroradiology*, pp. 809-816, 2010.

[79] A. Bruno, R. Collorec, J. Bezy-Wendling, P. Reuze, and Y. Rolland, "Texture Analysis in Medical Imaging," *Studies in Health Technology and Informatics* vol. 30, pp. 133-164, 1997.

[80] G. N. Srinivasan and G. Shobha, "Statistical Texture Analysis," *Proceedings of World Academy of Science, Engineering and Technology* vol. 36, pp. 1264-1269, 2008.

[81] R. L. Kashyap and R. Chellappa, "Estimation and Choice of Neighbors in Spatial-Interaction Models of Images," *IEEE Transactions on Information Theory* vol. 29, pp. 60-72, 1983.

[82] Siemens Healthcare, Pediatric MRI, [Online]. Available: `http://www.healthcare.siemens.com/magnetic-resonance-imaging/clinical-specialities/pediatric-mri/applications`

[83] V. Kumar, Y. Gu, S. Basu, A. Berglund, S.A. Eschrich, M.B. Schabath, K. Foster, H.J. Aerts, A. Dekker, D. Fenstermancher, D.B. Goldgof, L.O. Hall, P. Lambin, Y. Balagurunathan, R.A. Gatenby, R.J. Gillies, "Radiomics: the process and the challenges," *Magnetic Resonance Imaging,* vol. 30, no. 9, pp. 1234-48, 2013.

[84] G. Reynolds,"Magnetic Resonance Spectroscopy Methods for Childhood Brain Tumours," PhD Thesis, University of Birmingham, 2009.

[85] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine,* vol. 23, pp. 89-109, 2001.

[86] J. Rossiter,"Multimodal Intent Recognition for Natural Human-Robotic Interaction," PhD Thesis, University of Birmingham, 2011.

[87] S. Srihari, Center of Excellence for Document Analysis and Recognition, The State University of New York, Artificial Neural Networks [Online]. Available: `http://www.cedar.buffalo.edu/~srihari/CSE555/Chap6.Part1.pdf`

[88] C. Parmar, E. R. Velazquez, R. Leijenaar, M. Jermoumi, S. Carvalho, R. H. Mak, S. Mitra, B. Uma Shankar, R. Kikinis, B. Haibe-Kains, P. Lambin, H. J. W. L. Aerts, "Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation," *PLoS One,* 9:e102107, 2014.

[89] C. A. Graham, T. F. Cloughesy, "Brain tumour treatment: chemotherapy and other new developments," *Seminars in oncology nursing,* pp. 260-272, 2004.

[90] K. H. Peacock, G. J. Lesser, "Current therapeutic approaches in patients with brain metastases," *Current treatment options in oncology,* vol. 7, pp. 479-489, 2006.

[91] M. Grech-Sollars, "Diffusion MRI for Characterising Childhood Brain Tumours," PhD Thesis, University College London, 2014.

[92] V. T. DeVita, S. Hellman, S. A. Rosenberg *Cancer: principles and practcices of oncology,,* Lippincott Williams and Wilkins, 7th Edition, Philadelphia, 2014.

[93] T. E. Merchant, I. F. Pollack, J. S. Loeffler, "Brain tumours across the age spectrum: biology, therapy and late effects," *Seminars in radiation oncology,* pp. 58-66, 2010.

[94] D. J. Kroon, Snake: Active Contour - File Exchange - MATLAB Central [Online]. Available: `http://uk.mathworks.com/matlabcentral/fileexchange/28149-snake---active-contour/content/Snake2D.m`

[95] L. Harrison, P. Dastidar, H. Eskola, R. Jarvenpaa, H. Pertovaara, T. Luukkaala, R.-L.Kellokumu-Lehtinen, S. Soimakallio, "Texture analysis of MRI images of non-Hodjkin lymphoma," *Computers in Biology and Medicine,* vol. 38, pp. 519-524, 2008.

[96] L. Harrison, R. Nikander, M. Sikio, T. Luukkaala, M. T. Helminen, P. Ryymin, S. Soimakallio, H. J. Eskola, P. Dastidar, H. Sievanen, "MRI texture analysis of femoral neck: Detection of exercise load-associated differences in trabecular bone," *Journal of Magnetic Resonance Imaging,* vol. 34, pp. 1359-1366, 2011.

[97] L. Harrison,T. Luukkaala, H. Pertovaara, T. O. Saarinen, T. T. Heinonen, R. Jarvenpaa, S. Soimakallio, P.-L. Kellokumpu-Lehtinen, H. J. Eskola, P. Dastidar, "Non-Hodgkin lymphoma response evaluation with MRI texture classification," *Journal of Experimental and Clinical Cancer Research,* vol. pp. 28-87, 2009.

[98] A. Larroza, D. Moratal, A. Paredes-Sanchez, E. Soria-Olivas, M.L.Chust, L. A. Arribas, E. Arana, "Support vector machine classification of brain

metastasis and radiation necrosis based on texture analysis in MRI," *Journal of Magnetic Resonance Imaging,* 2015.

[99] K. P. Murphy, Naive Bayes Classifiers [Online]. Available: `http://www.ic.unicamp.br/~rocha/teaching/2011s1/mc906/aulas/naive-bayes.pdf`

[100] Leif E. Peterson, "K-nearest neighbor," *Scholarpedia,* vol. 4(2) pp. 1883, 2009.

[101] K. J. Geras, "Prediction Markets for Machine Learning," Masters Thesis, University of Warsaw, 2011.

[102] S. Sayad, Artificial Neural Networks [Online]. Available: `http://www.saedsayad.com/artificial_neural_network.htm`

[103] G. Pandey, A. Dukkipati, "Minimum Description Length Principle for Maximum Entropy Model Selection," *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on,* pp. 1521-1521, 2013.

[104] J. Rissanen, "Modeling by shortest data description," *Automatica,*vol. 14, no.5, pp. 465-658, 1978.

[105] E. H. Moore, "On the Reciprocal of the General Algebraic Matrix," *Bulletin of the American Mathematical Society,*vol. 26, pp. 394-395, 1920.

[106] K. R. Gray, "Machine Learning for Image-based Classification of Alzheimer's Disease," PhD Thesis, Imperial College London, 2012.

[107] L. Yu, H. Liu, "Feature Selection for High-Dimensional Data: a Fast Correlation-based Filter Solution," *Proceedings of the Twentieth International Conference on Machine Learning,* 2003.

[108] C. D. Manning, P. Raghavan, H. Schutze, "The bias-variance tradeoff," Introduction to Information Retrieval, Cambridge University Press. 2008.

[109] K. P. Murphy, Machine Learning: A Probabilistic Persoective, The MIT Press, Cambridge, Massachusetts; London, England, 2012.

[110] S. Tantisatirapong,"Texture Analysis of Multimodal Magnetic Resonance Images in Support of Diagnostic Classification of Childhood Brain Tumours," PhD Thesis, University of Birmingham, 2014.

[111] E. Ben Othmen, M. Sayadi and F. Fniaech, "3D Gray Level Co-Occurence Matrices for Volumetric Texture Classification," *Proceedings of the 3rd International Conference on Systems and Control,* 2013.

[112] D.H. Xu, A. S. Kurani, J. D. Furst, D. S. Raicu, "Run Length Encoding for Volumetric Texture," *Visualization, Imaging and Image Processing,* 2004.

[113] A. A. Brandes, E. Franceschi, "Genetic variation in pediatric and adult brain tumors," *Neuro-Oncology,* 2010.

Colophon

This document was typeset using LaTeX, in *Computer Modern Roman*, with a point size of 12 and double line spacing. As per the University of Warwick guidelines, a margin of 1.5 inches was kept on the left hand side. 1 inch were kept on the right hand side margin, and page numbers were typed at 1 inch into the page.

All the work presented in this thesis was carried out on Microsoft Windows 7 and OS X (initially Mavericks and later, Yosemite) platforms.

MR image inspection and slice selection were carried out using RadiANT DICOM viewer. Extraction of textural features was carried out using MaZda. Implementation of machine learning algorithms was done in python, using Orange library.

MATLAB was used to segment the images using the Snake GVF algorithm and to carry out KM survival analysis using the log-rank test.

Most plots were produced using MATLAB and Orange.